Concepts and concreteness in psycholinguistics

Lewis Pollock

UCL

PhD Linguistics

I, Lewis Martin Edward Pollock, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Lewis Pollock

August 2018

Abstract

Title: Concepts and concreteness in psycholinguistics

This thesis is about the concrete-abstract distinction ('concreteness') as it applies in psycholinguistic research and theories of concepts. Concreteness is one of the most-investigated psycholinguistic variables, and is also the basis for major disputes about the nature of the human conceptual system. However, I argue that concreteness is not actually a useful construct, and that the units of the conceptual system do not neatly match up with words of natural language, as is often assumed in the experimental and theoretical literatures.

I dispute evidence for 'concreteness effects', whereby words with high concreteness ratings exhibit processing differences relative to words with low concreteness ratings. The concreteness measure itself has statistical properties that invalidate it as a psycholinguistic tool. I report four new experiments designed to take into account these troublesome statistical properties, and maximise the chances of finding a concreteness effect. Counterintuitively, in three out of four experiments, the effect disappeared, and in the fourth it was extremely small. I suggest that evidence for concreteness effects is not as strong as it appears to be. Furthermore, even if the effects are real, current explanations of them still fail in various ways.

I also consider how the concrete-abstract distinction intersects with popular theories of concepts and cognition, with an emphasis on two in particular (a Fodorian Language of Thought, and a Barsalou-ian Simulator theory). Using the alleged 'abstract' concept JUSTICE as an example, I argue that from the point of view of these theories, some abstract concepts are explanatorily vacuous: they do not actually offer any insight into behaviour or cognition. I conclude that although many 'concrete' items belong in our theories of concepts, some alleged 'abstract' concepts aren't concepts at all. I explore some positive implications of this conclusion for theories of word meaning, and for theories of concepts in general.

Impact statement

The chief practical impact of this project concerns the psycholinguistic variable, concreteness. Concreteness is one of the most-investigated psycholinguistic variables, and literally hundreds of concreteness studies are published each year. Concreteness is not just used in psycholinguistics: it also features in computational linguistics research, it raises issues in the philosophy of mind, and it has even been suggested as a potential marker for a rare neurodegenerative disease. However, the principle conclusion of this project is that the way concreteness is operationalised in empirical work is fundamentally flawed, and that the measure suffers from a large degree of uninterpretable noise. Given the widespread use of concreteness measures, this is an extremely important issue, because a lot hangs on whether these measures work 'as advertised'. Although this thesis points out some problems with the concreteness measure, it also suggests a way around them. This aspect of the project has already been published as Pollock (2017).

Aside from the methodological validity of the concreteness measure itself, the other main outcome of this project relates to big-picture issues in cognitive science, philosophy of mind, and any number of related disciplines. There is a general model of human cognition that is very popular: psychologists, philosophers, and cognitive scientists tend to schematise cognitive processes as being made up of cohesive parts. This thesis argues that one assumption of this model is needlessly constraining, and that we can improve the model by relaxing it. The assumption is that words of natural language necessarily pick out elements of a theory of concepts. If this line of argument is correct, then it has positive implications for many research programs. For example, many challenges in cognitive science are typically attempted from within this general model: simulating cognitive processes, explaining how children learn, developing theories of utterance interpretation, and so on. My hope is that the conclusions of this thesis may make some of these goals more attainable by improving the framework with which we think about them.

Acknowledgements

The list of people I would like to thank for helping me with this project is too long to present in full here. But I would like to mention my wonderful parents, Kristian and Martin, for giving me every opportunity I could have asked for. Thank you to Joey for long phone call discussions; thank you to Winston for getting up at 5am on a Saturday; thank you to Liam for being good at booze. Thank you to all my friends, who did not mind when I was busy and were there when I was not.

The UCL Pragmatics Reading Group gave me a lot of important feedback over the years and shaped my thinking in many ways. Thanks to Tim Pritchard for his insight on some key issues.

On the practical side of things, I am indebted to Wing Yee Chow for help with the EEG data collection, and Matthew Jones for statistical advice. Thanks also to my two supervisors, Robyn Carston and Sebastian Crutch, who let me do basically whatever I wanted and took the result seriously every time. Last, but the opposite of least, thank you also to Danielle, without whom I would not be able to get up in the morning.

Table of contents

Table of figures

Chapter 1: Introduction

This thesis is about the concrete-abstract distinction ('concreteness') as it applies in psycholinguistic research and theories of concepts. Although, as we shall see, it has proven extremely difficult to come up with a rigorous definition of what the concrete-abstract distinction amounts to, concreteness is almost universally believed to be a central issue in work that investigates the human conceptual system. The idea is that there is a fundamental ontological distinction instantiated in the mind-brain between 'concrete' (sensorimotorically-derived) concepts, and 'abstract' (non-sensorimotorically-derived) concepts. In this thesis I have two aims. My first aim is to show that this belief is simply mistaken, and that actually concreteness is not a useful construct in either empirical psycholinguistic work, or in theories of cognition and conceptual processing. To do this, I will draw on data from a wide range of psycholinguistic experiments, the results of my own replication experiments, and also analyses of some popular theories of concepts and general cognition. My second aim is, having rejected the utility of concreteness as both an experimental and theoretical tool, to replace the concrete-abstract distinction with a more useful way of thinking about concepts. Work on concreteness often starts with an assumption about the relationship between the concepts we have and the words we know, which we can phrase in various ways. We might say that 'words encode concepts', or that 'word meanings are concepts', and so on. This new way of thinking about concepts is likely to seem unpalatable at first because it involves rejecting this seemingly reasonable assumption.

That being said, I believe there are substantial benefits to the view that I am going to try and motivate in the following chapters. Currently, almost everyone accepts that our theories of abstract conceptual content are severely impoverished relative to our theories of concrete conceptual content. As we shall see, this generates huge problems for otherwise-promising accounts of cognition, and has far-reaching implications for other areas of research, such as theories of communication that hold that word meanings are concepts. The most substantial benefit of the position I advocate is that it provides a way of avoiding these problems altogether, while maintaining the considerable explanatory power of these theories. The view that I am going to try and convince you of has two components. First, many of the words we know simply do not pick out basic elements of our theories of concepts. Second, in

the general case this phenomenon occurs with alleged word-concept pairings which received wisdom tells us are 'abstract'. My solution to the general problems concerning abstract conceptual content might be thought of as a methodology for determining what concepts there are. If an alleged concept is explanatorily vacuous and impossible to incorporate into a *theory* of concepts, then it is a bad candidate for concept-hood, and we should not *assume* that it belongs in our theory. If a problematic 'abstract' concept turns out not to be a concept after all, then our theories of concepts and cognition have one less problem to deal with for every alleged concept that fails this test.

I appreciate that at the moment this might seem difficult to swallow, given the fundamental importance placed on the concrete-abstract distinction and the wealth of experimental data on Concreteness effects. So now I will give an overview of the structure of the thesis in such a way that (hopefully) makes it clear how I intend to arrive at this conclusion. In Chapter 2, I will give an overview of concreteness as an experimental tool; that is, as a psycholinguistic variable. I will explain why concreteness is held to be so important, and provide a necessarily partial summary of the massive array of experimental findings that speak to this importance. Over more than half a century, independent teams of researchers have obtained experimental evidence to the effect that words with high concreteness ratings exhibit processing differences relative to words with low concreteness ratings. These experimental differences are taken to index some fundamental properties of the structure of the human conceptual system and the kinds of mental representations which constitute its resources. So I will also consider theoretical explanations of these experimental findings, focusing in particular on Lexical Decision data as an illuminating case study. In Chapter 2 and later in Chapters 5 and 6, where I consider the list memory and EEG literatures in greater detail, I argue that these theoretical explanations of the data all fail in various ways. I end Chapter 2 with one relatively self-contained argument against the validity of the concrete-abstract distinction in a particular area. This argument is a response to Kousta et al.'s (2011) recent proposals regarding a special role for 'emotional' information (information captured from experience of affective states) in individuating abstract concepts. I think that Kousta et al. are more than likely correct to suggest that this kind of experience is crucial for individuating our emotion concepts and instantiating them in cognitive processes. However, if that is the case, then we have no grounds for calling these emotion concepts 'abstract': from the point of view of theories of acquisition and representation, they do not have properties that are interestingly different from the

properties of concepts we assume to be concrete. Therefore, in this particular instance, we can do away with the concrete-abstract distinction.

Next, in Chapter 3, I begin to set out my arguments against the assumption that words and concepts stand in a reliable correspondence with one another. I consider how the concrete-abstract distinction intersects with popular theories of concepts and cognition, focusing particularly on a Fodorian language of thought (Fodor, 1998, 1975), and a Barsalou-ian simulator theory (Barsalou, 1999; Barsalou et al., 2008). I show that, although philosophers and psychologists disagree about the details of what a theory of concepts should look like, there is actually relatively widespread agreement about two absolutely non-trivial issues. Firstly, theorists tend to agree that a productive way to model human thought is to imagine it as being made up out of cohesive parts. We assume that these parts have properties that allow them to play certain roles in cognitive processes, such as categorisation and/or thinking about an entity 'as such'. The term "concept" refers to whatever these parts turn out to be. Secondly, as mentioned above, theorists also tend to agree that the words of a natural language stand in a reliable correspondence with these parts. So, however many concepts we have, we will at least have a DOG concept; a MOON concept; a JUSTICE concept, and so on.

For the remainder of Chapter 3, I argue that if we accept the first proposal, namely that concepts are parts of thoughts, then we should reject the second proposal, namely that there must be a concept for every word we know. I will show that no matter whether you believe the powers of the mind should be attributed to a language of thought, or to a network of simulators and frames, every theory works relatively well when it comes to concrete concepts. However, as others have pointed out, these theories tend to fall apart when it comes to abstract concepts. Using the alleged concept JUSTICE, an abstract concept par excellence, I show that the reason for this could simply be that these problematic abstract concepts are explanatorily vacuous. From the point of view of any given theory, JUSTICE does not actually explain any human behaviour or cognition. A theory of cognition can get by perfectly well without positing JUSTICE. Concepts are posits that we want to use to explain behaviour and cognition, and we should not posit concepts with no explanatory value, just as no other scientific theory should entertain posits with no explanatory value. I conclude that JUSTICE isn't a concept at all, because it is not useful to include it in our theory of concepts *however that theory looks.* I suggest that the same strategy that eliminated JUSTICE may be used to eliminate other troublesome abstract concepts as well. The upshot of this is that the concrete-abstract distinction collapses:

perhaps the only distinction to be drawn is between concepts that do happen to naturally match up with a word of natural language, and those that don't. I end Chapter 3 by acknowledging two major objections to this conclusion. Objection 1 is that there *must* be something psychologically and theoretically relevant about the concrete-abstract distinction because otherwise we would have no explanation of the vast quantity of experimental effects discussed in Chapter 2. Objection 2 is that there *must* be a concept for every word we know, because every psycholinguistic study and many popular theories of communication hold that word meanings are concepts: if there is no such thing as JUSTICE, then what is the meaning of the word, 'justice'?

I spend the remainder of the thesis outlining my responses to these two objections. First, I tackle the objection regarding experimental concreteness effects. In Chapter 4, through a simple analysis of a relatively new 40,000 word concreteness norm database (Brysbaert et al., 2013), I show that concreteness as a psycholinguistic variable suffers from a huge statistical anomaly. Words with mean values located at the extreme ends of the scale have accurate concreteness scores that genuinely track participants' judgements. However, words with mean values located in the middle of scale do not have accurate concreteness scores: their mean value is an illusion created by taking the average value of noise. In reality, participants were disagreeing about how to rate these words, and were not using concreteness as a linear scale, as it is often assumed to be. Even worse, I show that the stimuli featured in the 'abstract' condition of every concreteness experiment whose stimulus list I could obtain did not actually come from the truly abstract end of the concreteness scale. Instead, they tended to come from the middle of the scale, where the concreteness measure is just noise. This makes the experimental literature extremely difficult to interpret, because we have not actually been comparing responses to concrete words with responses to abstract words. Instead, we have been comparing responses to words that participants agree about (which were also 'concrete') with responses to words about which they disagree (and therefore have no grounds for calling 'abstract').

In Chapter 5, I report three list memory replication experiments designed to address the statistical problems with concreteness outlined in Chapter 4. I chose list memory paradigms because concreteness was originally used as an explanation of list memory paradigm effects; these experiments formed the evidence base for Paivio's (1991, 1986) classic Dual Coding Theory (DCT); and because list memory concreteness effects are relatively consistent in comparison to other experimental paradigms. In these new experiments, the contrast between concrete and abstract

conditions was maximised, and only words for which the concreteness measure is actually valid were featured in those conditions. In one experiment, marginal evidence in favour of the null hypothesis of no concreteness effect was obtained. In the second experiment, the statistical analyses were inconclusive. In the third experiment, a small concreteness effect was obtained. Note that these are surprising findings, given that the chances of finding a concreteness effect and the magnitude of these effects should have been maximised.

In Chapter 6, I report a sentence processing EEG experiment that measured responses to concrete and abstract target words. Concreteness effects in EEG are especially interesting because historically, behavioural responses (concrete words easier to process than abstract words) have not matched up with the ERPs produced (concrete words elicit larger N400s than abstract words, and the amplitude of the N400 is typically correlated with processing *difficulty*). As with the list memory experiments reported in Chapter 5, the concreteness measures of the stimuli featured in this experiment were controlled so that they were accurate reflections of participants' judgements, and the contrast between conditions was maximised. In spite of this, reasonably strong evidence in favour of the null hypothesis was obtained. I close this chapter with alternative explanations for why it is that previous EEG studies have found concreteness effects.

In Chapter 7, I draw together what I take Chapters 4-6 to show. They show that, contrary to what is generally assumed, actually concreteness effects are extremely fragile. Counterintuitively, if we maximise the contrast between concrete and abstract conditions, the effect tends to disappear. There are alternative explanations of reported concreteness effects that do not involve concreteness at all, but instead appeal to less controversial and less theoretically problematic mechanisms for driving differences across conditions. Even if we accept that concreteness effects are 'real', we still have to accept that there are huge problems with the way we currently operationalise it in experiments, given the statistical problems outlined in Chapter 4. We would also still be in the uncomfortable position of having to explain why differences between responses to concrete stimuli and responses to stimuli for which the measure is uninterpretable should be thought of as *concreteness* effects. I conclude that although Objection 1 has not been definitively resolved, a lot more work has to be done in order to show that our experiments really do produce concreteness effects, and therefore at the very least Objection 1 is not fatal to the view I am trying to convince you of.

In Chapter 8, I respond to Objection 2: if there is no such thing as the abstract concept JUSTICE (and indeed, some other 'abstract' concepts), then how are we to explain what the meaning of the word 'justice' is? My response to this objection is to point out that theories of meaning and of communicative success tend to assume that word meanings and concepts are the same thing, but that this really is just an assumption. There are alternative ways to account for the relationships between thought, language, and communicative success that do not require us to hold that word meanings are concepts. Instead of conceiving of communicative success as being the result of a compositional operation on concepts at the individual word level, we can instead account for communicative success by assuming that agents can store the inferences and conclusions they drew in responses to past usages of words. I argue that not only can this past usage information plausibly bear the same load that concepts are currently assumed to bear, but that actually in some ways a past usage account is superior to a conceptual account. Relevance Theory (Sperber and Wilson, 1998, 1995) is an extremely influential and powerful theory of communication that holds that word meanings are concepts. I show how Relevance Theory can be adapted in a relatively straightforward way so that a *necessary*, as opposed to *contingent,* link between words and concepts is severed. In this way, I hope to have offered a way of dealing with objection 2. In Chapter 9, I draw the various strands of my arguments together, and sketch out how we might treat concepts going forward.

Chapter 2: Concreteness in psycholinguistic experiments

Concreteness has become one of the most-studied variables in the psycholinguistic literature. Since Paivio et al. (1968) published one of the first large-scale databases of word concreteness norms, so-called 'concreteness effects' have emerged in a variety of investigations of many different cognitive processes. There are hundreds of experimental reports that purport to show that words with high concreteness ratings exhibit processing differences relative to words with low concreteness ratings. In this chapter, I will give an overview of what concreteness is, how it is operationalised in psycholinguistic experiments, and what role it plays in theories of cognition. I will offer a necessarily partial survey of the experimental literature on concreteness effects, although I will consider a range of behavioural and neuroimaging paradigms conducted over a period of many decades.

In this chapter, I pay special attention to lexical decision studies because of what I believe to be a curious gap between data and theory with respect to this paradigm (in Chapters 5 and 6, I focus on list memory and EEG experiments in more detail). Although recent experimental reports on concreteness effects frequently begin with an introduction to the effect that concrete words elicit faster reaction times than abstract words in lexical decision, a consideration of the entire lexical decision literature and a close look at the studies that are often cited reveals that this simply isn't true. Furthermore, Dual Coding Theory (Paivio, 1991, 1986) was often used in order to explain these effects, even though Paivio himself stressed that Dual Coding Theory as he conceived of it does not make predictions about lexical decision experiments (Paivio, 2013). Perversely, even if we did have strong evidence for concreteness effects in lexical decision, we currently have no way of explaining why these effects occur; and the theory that is most tested with respect to these effects does not actually predict them. As we shall see in Chapters 5 and 6, this kind of issue comes up in relation to other experimental paradigms as well. I end this chapter with one relatively self-contained argument against the validity of the concrete-abstract distinction, which is a response to Kousta et al.'s (2011) recent proposals regarding an important link between what they call 'emotional' information, and abstract concepts. I argue that if we accept their proposals (and I agree that they are

plausible), then we have no grounds for calling emotion concepts ‘abstract’ in the first place. This is because we can tell the same kind of causal story about the acquisition and instantiation of a standard concrete concept as we can of an emotion concept as we can of any standard concrete concept.

2.1 Overview of concreteness

First, let us consider what a word’s concreteness rating actually is, how it is derived, and what implications the word concreteness measure has for theories of the structure of the human conceptual system. A word’s concreteness rating is derived by simply asking a group of participants, typically numbering between twenty and thirty, to rate that word for concreteness on a Likert scale. A low score indicates that a word is highly ‘abstract’, whereas a high rating indicates that a word is highly ‘concrete’. The norming instructions that were given to participants in the original Paivio et al. (1968) study are reproduced below. These instructions provide an operational definition of ‘concreteness’ and ‘abstractness’ for participants:

> Nouns may refer to persons, places and things that can be seen, heard, felt, smelled or tasted or to more abstract concepts that cannot be experienced by our senses. The purpose of this experiment is to rate a list of words with respect to "concreteness" in terms of sense-experience. Any word that refers to objects, materials or persons should receive a high concreteness rating; any word that refers to an abstract concept that cannot be experienced by the senses should receive a low concreteness rating. Think of the words "chair" and "independence." "Chair" can be experienced by our senses and therefore should be rated as high concrete; "independence" cannot be experienced by the senses as such and therefore should be rated as low concrete (or abstract).

Spreen and Schulz (1966, p. 460), reused in Paivio et al. (1968)

This operational definition of concreteness has not changed very much over the years, although different norming databases do feature different instructions. In Brysbaert et al.’s (2013) recently published 40,000 word concreteness norm database, the guidelines given to participants were as follows:

> A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e.g. to explain 'sweet ' you could have someone eat sugar; to explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; to explain 'couch', you could point to a couch or show a picture of a couch). An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning).

Note that, although in the more recent Brysbaert et al. instructions, some characterisation of abstractness is given ('[an abstract word's] meaning depends on language'), in both of these passages, abstractness itself is barely defined at all for participants. A concrete word gives us 'immediate (sense) experience' and 'refers to objects', whereas an abstract word is more or less just "any word that *isn't* concrete". That is, the definition of abstractness is essentially negative. Others have pointed this phenomenon out (Crutch and Ridgway, 2012; Hamilton and Coslett, 2008; Wiemer-Hastings and Xu, 2005), but it seems to me that concreteness researchers see it as a challenge that will be solved eventually without too much trouble, rather than a fundamental problem that would have to be dealt with before any theoretical explanation of concreteness effects could possibly work. I will return to this issue throughout this thesis, but for now I simply want to point out that this really is a remarkable state of affairs. For more than 50 years, we appear to have been able to design experiments that show that stimuli with property X exhibit processing differences relative to stimuli with property Y. The problem is that we can't actually say what property Y is. It seems strange to attribute fundamental structural properties to the mind-brain based on a distinction that has simply never been clearly drawn in the experimental literature.

In any case, the mean value of all participants' ratings is taken to be an accurate approximation of a word's position on an abstract-concrete continuum. The existence of concreteness effects, whereby words with high scores exhibit processing differences relative to words with low scores, is taken to be evidence that this abstract-concrete continuum is neuropsychologically instantiated. By assumption, words and concepts are closely linked such that the effects produced in concreteness experiments that feature words are taken to be indicative of how *conceptual* processes work, and what *conceptual* mental representations are like. So, if a participant encounters the word "canary" in an experimental task, then their response to this word tells us something about the concept, CANARY. Or, perhaps more realistically, if we compare responses to a group of concrete words to responses to a group of abstract words, we might be able to make inferences about the properties of and differences between concrete and abstract concepts in general. To see that this assumption really is very widely held in the psycholinguistic literature, we only need to look at any of the discussion sections of any of the experimental reports featured in this chapter. For example, one of Kousta et al.'s (2011, p. 14) stated aims is to show how '[emotional information]… contributes to word representation and processing, particularly for abstract concepts'. They attempt to show this with evidence provided by a lexical decision task, in which all a participant has to do is make decisions about letter strings presented to them on a computer screen. Barber et al. (2013) use the same methodology and talk of their single-word stimuli as 'referring to concepts'. Barber et al. use their lexical decision data to draw conclusions about the nature of the mental representations involved in recognising words in the task, and in their discussion section they explicitly take these mental representations to be conceptual in nature. Likewise, Binder et al. (2005) also take lexical decision data to be informative about concepts.

The idea that recognising or processing a word automatically "activates" a concept that corresponds to that word is common throughout the psychological literature, and is not just found in lexical decision studies. For example, Geng and Schnur (2015) report an odd-one-out identification task, Romani et al. (2008) report a list memory task, de Groot (1989) reports a word association task, and so on, and all of these researchers also draw conclusions about conceptual processing or structure on the basis of their data. For the moment, I am not arguing that this approach is flawed, although that will be one of the conclusions I arrive at by the end of this thesis. For now, I am simply pointing out the close connection between single

lexemes of English and concepts that is assumed throughout the psycholinguistic literature that observes the concrete-abstract distinction.

2.2 Evidence and explanations of concreteness effects

If there are reliable behavioural and neuroimaging effects that can be attributed only to the manipulation of word concreteness as it is operationalised in traditional concreteness norms, then there needs to be a neuropsychological model that explains why these effects obtain. So one of the main tasks facing psycholinguistic theories of conceptual representation is the construction of an account of how it is that concrete and abstract concepts differ in terms of their informational content and structure, such that this difference would produce concreteness effects. Generally, theorists propose that concrete and abstract concepts are 'represented differently'. That is to say, the structure, the content, and/or neural architecture that supports concrete items in the conceptual system are qualitatively and/or quantitatively different to those that support abstract items. Although I will end up arguing that the concreteness measure might not actually index anything psychologically relevant, I do want to acknowledge that prima facie it does seem like something that might do so. If there are properties that define a cognitively relevant ontology of concepts, concreteness seems like a good candidate: if for the time being we accept that words and concepts are closely aligned, it seems intuitively highly plausible that there is something about what constitutes a concept of 'elephants' (highly concrete) that is different to what constitutes a concept of 'paradoxes' (highly abstract).

So, what kind of evidence for concreteness effects has been obtained, and what have theorists tried to infer about the conceptual system on the basis of this evidence? A complete list of concreteness effects and their theoretical explanations might be thousands of articles long. But here is a representative sample of the work that has been done. Initially, concreteness effects were obtained in paired associate recall paradigms in which participants were presented with sequences of pairs of words or phrases (Begg, 1972; Nelson and Schreiber, 1992; Paivio et al., 2000, 1994). After a sequence of pairs had been presented, participants were required to recall as many words as they could from the sequence. The precise task demands could vary somewhat. In some experiments, participants were given no cues, and simply asked to write down or repeat whatever words they could recall. In other experiments, participants were provided with the first word of each pair during the test

phase, which functioned as a cue. They then only had to try and recall the other member of each pair. The relationships between the words in each pair were also manipulated in various ways. Both words in a pair could be concrete; or both could be abstract; or the cue could be concrete with an abstract target; or vice versa; or they could be semantically related; and so on. A consistent finding in these experiments was that participants were better at remembering items from concrete word pairs than they were at remembering items from abstract word pairs.

Paivio (1991, 1986) developed his highly influential Dual Coding Theory partially as an explanation for why these memory concreteness effects occurred. In brief, DCT posits two types of representational unit: imagens and logogens. Each type of representational unit features in different classes of cognitive process, called 'codes' in the context of the theory. There is a 'verbal' code, and a 'non-verbal' code. The verbal code operates over 'logogens', whereas the non-verbal code operates over 'imagens'. Imagens and logogens are specialised for the storage and processing of two different types of perceptually-derived information, and as a result they have different properties and relational structures. Logogens are completely content-less: they are simply sensorimotor recordings of hearing or seeing a word, and they function like labels. Logogens have associative connections with other logogens, as well as referential connections with imagens, which are stored mental representations of sensorimotor experience of objects and events. In DCT, a word (i.e. a content-less logogen) is concrete to the extent that it has referential connections with content-*ful* imagens. A word (logogen) is abstract to the extent that it lacks such connections, and is only associatively connected to other content-less logogens. In DCT, a referential connection is any connection between the representations of different types (i.e. between an imagen and a logogen, as opposed to between two logogens). The reason that concrete words are easier to recall than abstract words is because they allow participants to employ a strategy that helps with recall: a participant can leverage the imagens that are referentially connected with a concrete logogen in order to generate mental imagery. This mental imagery leaves a relatively large cognitive footprint, which makes the logogens that ultimately triggered the imagery easier to recall. In the case of abstract logogens, there are less referentially-connected imagens on which to deploy this strategy, and so they are harder to recall.

I should note here that although this formulation of DCT is completely explicit throughout all of Paivio's work (2013, 1991, 1986), often his theory is misinterpreted on two fronts. Firstly, many researchers seemed to have assumed that DCT is committed to the claim that concrete words should always enjoy a processing

advantage over abstract words no matter what experimental task is involved. For example, in the lexical decision literature we shall shortly examine in some detail, it is often claimed that DCT predicts that concrete words should be recognised faster than abstract words (Barber et al., 2013; Kousta et al., 2011; Kroll and Merves, 1985). However, Paivio denies this and has quite consistently claimed that DCT predicts that concreteness effects should only occur when a task is such that a processing strategy on the part of the participant could make use of the structural properties of imagens and logogens (Paivio, 2013). Secondly, some modern theorists take DCT as a starting point for their own accounts of cognition and concepts (Barsalou et al., 2008; Dove, 2011), but they seem to misunderstand how Paivio conceived of logogens and imagens. For example, Dove (2011) writes that:

> DCT claims that mental images are the basic constituents of the verbal and non-verbal systems… Perceptual symbols [Dove's preferred construct] differ from mental images in a number of important ways: for instance they need not be conscious, they can be schematic, and they are often multimodal.

But Paivio (1986) actually states that mental images are *not* the basic constituents of either system: mental images are *generated* from imagens under task constraints, but it is imagens themselves that the non-verbal code operates over. He also explicitly states that both verbal and non-verbal codes may operate over their respective kinds of mental representation in an unconscious, schematic, and multimodal way, and he emphasis that the term 'mental imagery' can apply to imagery generated from imagens of all modalities, and not just the visual modality. So it is unclear that the theory that Dove proposes is actually different to traditional DCT in this respect. In Barsalou et al.'s (2008) update of Barsalou's (1999) original Perceptual Symbol Systems theory, they state that '[DCT] assumes that deep conceptual processing occurs in both systems [i.e. both the verbal and non-verbal systems]'. Again, this seems somewhat at odds with what Paivio seemed to think: in Paivio and Sadoski (2011) he reiterated his view that logogens are 'relatively meaningless' and that the sole content-bearing mental representations in his theory were imagens.

There is a broad point to be made here, which is that it is hard to overstate the influence of DCT on modern theories of concepts and accounts of concreteness effects. The fundamental DCT idea that we can model the processes of the mind-brain as operating over different *kinds* of mental representation appears again and

again (Barber et al., 2013; Barsalou et al., 2008; Dove, 2011; Kousta et al., 2011). And although what a particular theorist means to include under the umbrellas of 'sensorimotor experience' and 'linguistic experience' will vary, it is still true to say that these two broad categories of mental representation are the same as the ones suggested by Paivio more than 50 years ago.

2.3 List memory paradigms

That issue aside, since these relatively early pair associate recall studies, a new series of list memory studies have also reported concreteness effects (Allen and Hulme, 2006; Miller and Roodenrys, 2009; Romani et al., 2008; Walker and Hulme, 1999). In these studies, participants were simply presented with lists of words, 5 to 8 words long. Each study contained multiple experiments on different sets of participants, so these four reports alone produced more than 10 different instances of concreteness effects. Each list was either composed entirely of concrete words, or entirely of abstract words. After each list had been presented, participants had to recall as many words as they could. Again, there were many different subtle modulations to the paradigm, such as distracting the participant with a concurrent task; requiring them to recall the list in a specific order; and so on, but the concreteness effect itself was more or less consistent across these manipulations. Participants were better able to recall concrete words than abstract words, unless the task demands were so high that performance was extremely poor for both types of stimulus. I return to a fuller discussion of these experiments in Chapter 5, but for now I note that in paradigms that require participants to recall lists of words that they have just been presented with, concreteness effects seem to be extremely prevalent.

2.4 fMRI paradigms

Concreteness effects have also been investigated extensively using both fMRI (Binder et al., 2005; Fiebach and Friederici, 2004; Jessen et al., 2000; Kiehl et al., 1999; Pexman et al., 2007; Skipper-Kallal et al., 2015) and EEG (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000). These methods from neuroscience are often combined with traditional behavioural measures such as sentence verification tasks and lexical decision, and the typical behavioural advantage for concrete words is often obtained (although Barber et al. (2013) is an exception to which I return when I discuss the lexical decision data). The typical aim of these studies is to assess patterns of brain activation elicited by target concrete and abstract words in order to determine whether there are statistically significant differences between these patterns. The thinking goes like this: Dual

Coding Theory (or some more recent development of it[1]) predicts that the patterns of brain activation produced in response to concrete words should be different to the patterns of brain activation produced in response to abstract words. Dual Coding Theory's reason for this prediction is that concrete words are referentially connected to more imagens than abstract words, and will 'trigger' or 'activate' these imagens. Abstract words tend not to be referentially connected to imagens to the same extent, and so will not trigger as many of these representations by default. This should show up in fMRI as topographical differences in the magnitude of the BOLD response to the target words across conditions. Or, in EEG studies, this should show up as an interaction between the magnitude of the ERP waveform at a given time-point and the locations of electrodes placed on the scalp. Partly in response to the rise of currently-popular Embodied Cognition frameworks, terminology has moved on so that, instead of talking of imagens and logogens, researchers talk of 'multimodal' representations and 'linguistic' representations. Concrete words are assumed to activate more multimodal representations (mental representations derived from sensorimotor experience) than abstract words, which tend to activate networks of linguistic representations. Either way, the rationale for predicting a concreteness effect in these neuroimaging studies is the same. A concreteness effect should manifest as a topographical difference in patterns of brain activation because of what Holcomb et al. (1994, p. 723) call the 'spatial distinctiveness principle', which is the assumption that 'two or more distinct cognitive systems in the brain will tend to be more spatially distinct within the brain than will a single cognitive system.' The assumption here is that whatever 'linguistic' representations are, they belong to a different 'cognitive system' to that of the multimodal sensorimotor representations.

So, what are the results of these neuroimaging studies? In the fMRI literature, it is clearly the case that statistically significantly different patterns of brain activation have been found between responses to concrete words and responses to abstract words. The problem, as some of these researchers have noted, is that the patterns themselves are different in nearly every experimental report. Part of this is almost certainly due to the fact that researchers used different tasks and paradigms, and so that explains some of this variability. However, if there really are two spatially distinct cognitive systems which house mental representations of different types, *and* the concrete-abstract distinction aligns with these two systems, then I do not think we

[1] I should stress here that not all of these researchers take themselves to be conducting tests of the original DCT as Paivio proposed it, although some do (e.g. Barber et al., 2013; Fiebach and Friederici, 2004; Kounios and Holcomb, 1994).

would expect variability of the magnitude which we do in fact find. Binder et al. (2005) report that both concrete and abstract words activated areas in the left temporal lobe. Concrete words elicited greater activation in the angular gyrus and dorsal prefrontal cortex bilaterally. Abstract words elicited greater activation in the left inferior frontal gyrus, and Binder et al. conclude that 'processing' of abstract concepts occurred 'almost exclusively' in the left hemisphere. Kiehl et al. (1999) report that both concrete and abstract words activated the bilateral fusiform gyrus, anterior cingulate, left middle temporal gyrus, right posterior superior temporal gyrus, and left and right inferior frontal gyrus. They also report that abstract words elicited greater activation in the right anterior temporal cortex, and interpret this as 'support for a right hemisphere neural pathway in the processing of abstract word representations'. Jessen et al. (2000) found greater activation for concrete words in the lower parietal lobes bilaterally, the left inferior frontal lobe, and in the precuneus. Pexman et al. (2007) et al. found 'more widespread cortical activation' for abstract words than concrete words, in areas covering temporal, parietal, and frontal cortices.

For any one of these studies, we can pick another that found a conflicting result. Binder et al. find that abstract words 'almost exclusively' activate areas in the left hemisphere, whereas Kiehl et al. interpret their results as showing that abstract words preferentially activate an area in the right hemisphere. Pexman et al. find that abstract words elicit greater activation in areas distributed across much of the cortex, whereas Jessen et al. find greater activation for concrete words in more areas than abstract words. Binder et al. highlight the role of the left inferior frontal gyrus for abstract word 'processing'. Kiehl et al. find that the same area is equally involved in both concrete and abstract conditions. Jessen et al. find that the *same* area is more activated for concrete words. This situation is markedly different to what we saw with the list memory paradigms. In those studies, the same concreteness effect was found repeatedly: concrete words are easier to remember than abstract words. However, for the fMRI data, we seem to get a different result every time. Note that this observation does not just apply to classic Dual Coding Theory's predictions, but to modern investigations of the concrete-abstract distinction as well. Classic Dual Coding Theory predicts a relatively stable difference in brain activation across conditions that is reliably localised to certain areas, and we certainly don't find that. But the trouble is that we don't seem to find *any* reliable difference between concrete and abstract conditions, and sometimes experiments produce results that are in direct conflict. If an experimental contrast does not produce reliable effects, then that suggests that any neuropsychological theory that is built on this contrast may be

incorrect. This is especially true when it comes to concreteness, because concreteness is the basis for claims about fundamental structural properties of the human mind-brain. If these claims were true, we would expect these fundamental structural properties to manifest themselves in such a way that it was possible to detect them consistently. I will end this discussion of concreteness fMRI studies with the observation that, although we may have found a way to generate p-values below 0.05 when it comes to comparing regions of brain activation, this does not mean that we have come up with evidence in favour of there being a consistent experimental effect. In Chapter 4, where I consider some problems with the concreteness measure itself, I offer a speculative explanation for why these variable patterns of brain activation have been obtained.

2.5 EEG paradigms

In the EEG literature, the picture is healthier, although the experimental results here are marked by a rather counterintuitive finding. EEG experiments record 'event related potentials' (ERPs) while participants are engaged in a task. ERPs are tiny voltage changes measured by electrodes placed over the scalp, time-locked to the presentation of a stimulus (that is, the 'event'). The result is a waveform, with time on the x-axis and amplitude on the y-axis. There will be such a waveform for every electrode placed on the scalp. A typical stimulus in linguistic tasks is a single word located in the centre of a computer screen. Over the last few decades, certain types of stimuli and experimental paradigms have been shown to produce consistent patterns of ERPs, called 'components'. In some cases, specific properties of stimuli have been shown to modulate these components in specific and predictable ways. Researchers try to isolate properties of stimuli that have reliable modulatory effects, and then they try to infer things about cognitive processes on the basis of the properties of the stimuli that have been so isolated.

In concreteness research, a particularly important component is the N400 (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000). This component is called the N400 because it is a negative-going deflection in voltage that peaks approximately 400ms after a stimulus has been presented. The N400 is important for a number of reasons. It was discovered by Kutas and Hillyard (1980), who showed that when participants read sentences presented one word at a time, semantically anomalous words ("I accidentally burnt the *socks*") elicited a large negative deflection peaking 400ms after the word had been presented, whereas semantically congruent words ("I accidentally burnt the *toast*") elicited a

smaller negative deflection. Note that the difference here is relative: all content words will produce an N400, but words with different properties will produce N400s of different relative amplitudes. In later work, it was shown that this modulation was not driven by physical characteristics of the stimuli, because N400s to words that are semantically congruent but appear in an unexpected colour are the same as those that appear in an expected colour. The N400 is therefore taken to be at least partly sensitive to the contribution a word makes to the interpretation of the sentence it appears in. A number of other modulators of the amplitude of the N400 have been found since. For example, words that are less frequent elicit larger N400s than words that are more frequent (Van Petten and Kutas, 1990). N400 amplitude is also sensitive to priming, such that primed stimuli elicit smaller N400s than un-primed stimuli (Kutas and Federmeier, 2000); and to cloze probability, such that words with high cloze probability elicit smaller N400s than words with low cloze probability (Kutas and Hillyard, 1984). Because the N400 appears across a wide range of tasks, and is modulated by a wide range of stimulus properties, it quite widely accepted that it is probably the result of more than one cognitive process (or neural generator).

Note how, in all of these cases, the N400 amplitude is correlated with how *difficult* a particular stimulus is to process. It is well known that low frequency words elicit slower reactions than high frequency words; that un-primed stimuli are generally processed slower than primed stimuli; that unexpected stimuli are harder to respond to than expected stimuli. However, the curious thing about concreteness EEG research is that it has been consistently found that concrete words elicit larger N400s than abstract words in a variety of language processing tasks (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000). This is curious because, as we saw in previous sections, the majority of the time it has been found (or, claimed) that concrete words are easier to process than abstract words: they are easier to remember and supposedly recognised faster in lexical decision. The theoretical interpretation of these behavioural findings is that there is something about how the human conceptual system is structured that makes it the case that concrete stimuli are easier to process than abstract stimuli in these experimental paradigms. This being so, we would expect the opposite result when it comes to the EEG data. We should expect N400s to abstract words to be larger than N400s to concrete words, because N400 amplitude seems to be correlated with processing difficulty, and most of the psycholinguistic literature suggests that abstract words are harder to process than concrete words. I will return to a fuller discussion of the EEG concreteness literature in Chapter 6, where I report my own concreteness

EEG experiment. For now though, it is fair to say that there are consistent concreteness effects in EEG paradigms: concrete words, somewhat paradoxically, elicit larger N400 amplitudes than abstract words.

2.6 Other paradigms

Before turning to a more in-depth discussion of the lexical decision data, I want to briefly stress that behavioural concreteness effects have turned up in a very large variety of paradigms, and are not just confined to lexical decision and list memory. Geng and Schnur (2015) report an odd-one-out decision task; Sabsevitz et al. (2005) report a semantic similarity judgement task; Sadoski et al. (1997) examined how well participants could define concrete words versus abstract words; Skipper-Kallal et al. (2015) simply asked participants to 'think deeply' about words while undergoing fMRI; Holcomb et al. (1999) report a word-by-word sentence congruency judgement task; and so on. Independent teams of researchers operating over a period of decades, using very different methods, have all found evidence that seems to point to the existence of concreteness effects. That is surely something that has to be explained, and the most natural explanation is that the concreteness measure that we used to obtain these effects really is indexing something psychologically important. Much of this thesis will be spent arguing that, unfortunately, this is incorrect, and that there are other reasons that these statistically significant results have been obtained. A lot of my empirically-grounded arguments will be based on an analysis of a huge concreteness norm database that appeared only recently (Brysbaert et al., 2013), and the people who conducted the studies cited in this chapter did not have access to this resource and so could not have raised the issues I am going to discuss.

2.7 Lexical decision paradigms

There are probably more lexical decision concreteness studies than any other kind. Here is a non-exhaustive list: Barber et al. (2013), Binder et al. (2005), Connell and Lynott (2012), Fiebach and Friederici (2004), James (1975), Kounios and Holcomb (1994), Kousta et al. (2011), Kroll and Merves (1985), Paivio and O'neill (1970), Richards (1976), Rubenstein et al. (1970), Winnick and Kressel (1965). Recent lexical decision studies often begin with the claim that, historically at least, concrete words elicit faster decision latencies than abstract words (Barber et al., 2013; Kousta et al., 2011), although Binder et al. (2005) is a notable exception. The question is: *why* is it that participants are faster to respond to words that label concrete objects than they are to respond to words that don't? This finding is especially

mysterious because you might think that lexical decision experiments would be unlikely to produce any concreteness effect at all. In lexical decision, all a participant has to do is verify whether isolated letter strings constitute words of English. As Paivio (2013) noted, Dual Coding Theory actually predicts that there should be no difference in decision latency between concrete and abstract words, because in order to verify either word type, a participant's cognitive system simply has to activate the appropriate logogen, and nothing else. The structural properties that differentiate concrete and abstract words in DCT are not relevant to the task, and so should not show up in it.

You also might think that more modern theories that couch explanations in terms of multimodal experience and linguistic representations would say much the same thing as DCT. However, modern theories tend to explain a decision latency advantage for concrete words via a mechanism that involves a facilitatory effect of semantic access that occurs to a greater degree for concrete words (e.g. Binder et al. (2005). The idea is roughly that encountering concrete words triggers by default more mental representations that constitute what we might call "the meaning" of those words, and that this activation feeds back to decision mechanisms somehow in such a way as to facilitate a choice about whether a letter string constitutes a real word. However, as Connell and Lynott (2012, p. 453) note, it is simply not true that a convincing majority of lexical decision experiments produce a decision latency advantage for concrete words. Consider Table 2-1 below, which is a reproduction of Gernsbacher's (1984) summary of relatively old lexical decision results:

Table 2-1 Reproduction of Gernsbacher's (1982) summary of lexical decision results

Study	Experiment	Result
(Winnick and Kressel, 1965)	n/a	Concrete worse than abstract
(Paivio and O'neill, 1970)	n/a	Concrete worse than abstract

(Richards, 1976)	1	Concrete better than abstract
(James, 1975)	1	Concrete better than abstract
	2	Concrete better than abstract
	3	Concrete equal to abstract
	4	Concrete equal to abstract
(Rubenstein et al., 1970)	n/a	Concrete equal to abstract
(Richards, 1976)	2	Concrete equal to abstract

In the experiments displayed here: nearly a third of the time response to concrete words is worse (slower; contained more errors) than abstract words, a third of the time response to concrete words is better than abstract words, and a third of the time there is no statistically significant difference between the conditions. As with the fMRI data, we can pick any study we like and find another that produced conflicting results. Gernsbacher (1984) reports her own lexical decision experiment and concludes that concreteness is not actually driving any performance differences across conditions. Instead, she claims that word familiarity is the real locus of the effect, and because familiarity had been frequently confounded with concreteness, this is the reason for the inconsistent results. The most recent concreteness lexical decision experiments now suggest that there is no difference between concrete and abstract decision latencies (Brysbaert et al., 2016), or that after partialling out the effects of other variables, there may be an advantage for abstract items (Barber et al., 2013; Kousta et al., 2011). If we place more credence on the results of these newer studies than the older studies (perhaps because of improvements in stimulus controls and statistical methodology), then we obviously should not endorse any theory that was designed to account for a decision latency advantage for concrete words over abstract words.

2.8 Discussion of Kousta et al. (2011)

So, in more recent studies that produced a latency advantage for abstract words, what is the explanation provided for this abstractness effect? I will consider Barber et al. (2013) in detail in Chapter 6, because their lexical decision results were acquired while they recorded EEG, and in Chapter 6 I report my own EEG experiment. For now I will focus on Kousta et al. (2011), who explore the hypothesis that the Emotional Valence associated with the referent of a word may be the crucial factor underlying a lexical decision advantage for abstract words over concrete words. This hypothesis was primarily motivated by two previous findings. Firstly, Altarriba et al.

(1999) found that emotion words (e.g. ‘anger’), concrete words, and abstract words all pattern differently when rated on various norming scales. This difference in rating patterns is taken to be potentially indicative of a psychologically relevant type distinction between emotion words and abstract words: emotionally valenced words tend to have low concreteness ratings but this might not make them ‘abstract’ in the same sense as non-emotionally valenced abstract words. Secondly, Kousta et al. (2009) found that words which participants associate any kind of emotion with, positive or negative, are responded to faster than emotionally neutral words in a lexical decision experiment. In Experiment 1, Kousta et al. (2011) found that abstract words elicited shorter decision latencies in a lexical decision task than concrete words. Taken together, these findings suggest the possibility that the reason that abstract words were responded to quicker than concrete words in Kousta et al.’s Experiment 1 might be because emotionally valenced words have an advantage over neutral words, rather than because of an abstractness advantage per se.

In their Experiment 2, Kousta et al. (2011) selected words with neutral emotional valence that ranged from highly abstract (e.g. *minute, number, joy*) to highly concrete (e.g. *office, guest, voice*). They reasoned that if the ‘abstractness’ effect found in their previous analyses was really an emotional valence effect, then this effect should disappear when emotional valence was held constant and neutral. Although the procedure of Experiment 2 was the same as that of Experiment 1, Kousta et al. use a different statistical analysis. In Experiment 1, Kousta et al. conducted an ANOVA that directly compared a concrete condition to an abstract condition. In Experiment 2, Kousta et al. employ a Multiple Linear Regression analysis (MLR). So, in experiment 2, the question is not whether abstract words are responded to faster or slower than concrete words across experimental conditions, but whether these variables turn out to have a statistically significant effect in a multiple linear regression model, when emotional valence is held constant and neutral and a host of other predictor variables are also included in the model. There could be an issue here with multicollinearity.

In brief, the problem of multicollinearity applies in regression analysis when two or more predictors are correlated with each other and with the dependent variable. As well as other variables known or suspected to impact lexical decision latencies, Kousta et al. enter both concreteness and imageability into the same models, and these two variables are highly correlated. One way of thinking about what MLR does is that it asks the following question: once all of the other predictor

variables in the model are known, what is the value of adding predictor variable X? It asks this question for every predictor variable included in the model. To relate this to the present situation, when we include imageability in the MLR analysis, we are asking: once I already know a word's frequency, age of acquisition, length, concreteness, etc…, what is the predictive value of knowing that word's imageability score? If two or more of the predictor variables are correlated with each other and with the dependent variable, this generates huge problems. Concreteness and imageability are so highly correlated that Kousta et al. (2011, p. 16) note that 'it is invariably assumed [in the literature] that the psycholinguistic constructs of concreteness and imageability tap into the same underlying theoretical construct'. If concreteness and imageability are included in the same model, then because they are correlated, their association with reaction time might well be masked. This is because the MLR is asking: 'once a word's imageability is known, what is the value of also knowing its concreteness?' and 'once a word's concreteness is known, what is the value of also knowing its imageability?'. The answer to these questions is 'not much', because concreteness and imageability are highly correlated, so they are not providing different information. It is possible that by excluding imageability from the analysis, the predictive value of concreteness would rise (or vice versa). Multicollinearity problems can quite routinely result in predictor variables being assigned coefficients of the wrong polarity, which is especially troubling given that Kousta et al. (2011) make arguments on the basis that imageability is negatively correlated with reaction time, whereas concreteness is positively correlated. These considerations should not serve to completely invalidate Kousta et al.'s findings. They should simply motivate caution in the interpretation of their models.

Statistical methodology aside, Kousta et al. (2011) find that neither concreteness nor Imageability predicted log reaction time when Emotional Valence was kept constant and neutral, which is exactly what their hypothesis predicts. Surprisingly, despite the fact that Emotional Valence was held nearly constant (values for all words were between 4.25 and 5.75 on a 1-9 point scale), Emotional Valence was still a statistically significant predictor of reaction time. The higher the Emotional Valence of a word, the faster participants tended to respond to it. This result is also in line with their hypothesis that Emotional Valence underlies the advantage for abstract words reported in Experiment 1 and the MLR analyses. Kousta et al. report a final lexical decision experiment, Experiment 3, in which concreteness, Imageability, and Emotional Valence of words all varied across their respective scales. The procedure was the same as that of the previous lexical decision experiments, and the

data were analysed with MLR. Kousta et al. (2011) reasoned that if their hypothesis is correct, then in an MLR of the data from this experiment, concreteness and Imageability should not be statistically significant predictors of reaction time when Emotional Valence is included in the model. However, when Emotional Valence is removed from the model, concreteness and Imageability should be statistically significant predictors. This is what one would expect if the reason for the advantage for abstract words is because abstract words tend to have higher Emotional Valence than concrete words, and Emotional Valence is the real reason for the effect. Once again, this is exactly what Kousta et al. (2011) find.

To summarise, in Experiment 1 and Regression Analyses 1 and 2, Kousta et al. (2011) find a processing advantage for abstract words over concrete words in that abstract words were responded to faster in a lexical decision paradigm. In Experiments 2 and 3, Kousta et al. (2011) tested the hypothesis that the reason for this abstract word advantage was not due to abstractness per se, but due to an advantage for Emotionally Valenced words over Emotionally Neutral words. They tested this hypothesis by running lexical decision experiments in which Emotional Valence was either held constant, varied and included in an MLR model, or removed from an MLR model, and assessing the statistical significance of concreteness and Imageability in each model. They take their results to strongly support their hypothesis. In a model that includes Emotional Valence, concreteness and Imageability are not statistically significant predictors. In a model with Emotional Valence removed, concreteness is statistically significant, and the higher the concreteness, the slower participants are to respond. Ultimately, the extent to which one would want to rely on these findings will be a function of the extent to which one accepts MLR as a suitable tool for confirming experimental hypotheses. Although Kousta et al. (2011) provide an impressive array of converging results and meticulously controlled experimental stimuli, these undeniable strengths are tempered somewhat by the controversy surrounding MLR analysis in general (Flom and Cassell, 2007; Hurvich and Tsai, 1990; Ryan, 2008, p. 269; Whittingham et al., 2006). Putting this issue to one side, we shall now turn to the theory of abstract conceptual content that Kousta et al. (2011) advance on the basis of these results.

Kousta et al. (2011) identify their theory as being part of the ‘embodied cognition’ framework. Embodied cognition theorists hold that the neural systems that support higher level cognitive processes such as conscious thought are the same as (or significantly overlap with) the neural systems that support perception and action. Often, embodied cognition theorists talk of cognition being ‘grounded in’ perception

and action. Embodied cognition is generally taken to be in opposition to so-called 'amodal' theories of cognition. In amodal theories of cognition, the processes that realise cognition and the information that these processes operate on are fundamentally distinct from the processes that support perception and action, and the information derived from the sensorimotor systems. For the amodalist, cognitive phenomena are thought of as resulting from mental processes often characterised as functions. These functions are expressions in a language-like system that operates symbolically, somewhat like a computer language, or a logical calculus. For the amodalist, the mind is instantiated by the brain, but this might not be the only way of instantiating a mind. If we had total knowledge of the functions that govern cognition, it might not be inconceivable in principle for scientists of the future to build an artificial mind (although this might turn out to be impossible in practice). On the other hand, for embodied theorists this task would be harder, and perhaps impossible in principle. The embodied theorist holds that cognition is fundamentally intertwined with perception and action, so a *human-like* mind would need to be realised in a body able to sense and physically act like a human. A human mind could not be instantiated simply by some very sophisticated software running on some very sophisticated hardware, unless that configuration was somehow human-like in just the right way (note that this leaves open the possibility that there may be lots of other ways of instantiating something which we would want to call a mind; it just wouldn't be a human-like mind). I absolutely do not want to claim that all embodied theorists and amodal theorists would accept this conception of the debate: it is simply one way of thinking about the issues that are at stake.

As Kousta et al. (2011, p. 24) note, embodied theories have often been criticised on the grounds that they cannot possibly provide an account of abstract conceptual content (Goldinger et al., 2016; Mahon and Caramazza, 2008). Putting aside the issue that no one seems to be quite clear on exactly what the properties of abstract concepts are, the criticism is that if cognitive processes and the information that features in these processes are all fundamentally related in type to sensorimotor processes and information, then it is unclear how we come to be able to think and talk about abstract entities. This is because, despite the confusion just noted, one of the characteristics generally attributed to the things we think about using our abstract concepts is that we cannot have sensorimotor experience of the categories they apply to. If we arguably cannot derive our abstract concepts from sensorimotor experience, and if cognition is 'grounded in' sensorimotor experience, then how is it possible for concepts of paradoxes, injustices, or theories to feature in our mental lives? Kousta

et al. aim to provide a solution to this very important problem with their new theory. They propose that affective information (information pertaining to emotional states) plays a key role in the 'representation of' abstract concepts. They do not make it explicit what they mean by the phrase 'representation of', but I take them to mean that affective information is (partly) constitutive of the content of abstract concepts. This is supposed to get around the general problem of abstraction that faces embodied cognition theories: we can acquire and think with abstract concepts because their content is at least partially constituted by affective information derived from introspective bodily states. This information is not fundamentally different in type from the information available during actual experience of an emotional state. In this way, Kousta et al. hope to maintain the core tenets of embodied cognition while providing an account of how abstract concepts are acquired and realised in cognitive processes. Kousta et al. (2011) take their experimental results to strongly support this theoretical claim. Their thinking is that a lexical decision reaction time advantage attributable to the emotional valence associated with a word is evidence that affective information is constitutive of the content of the concept that the word is (generally) associated with.

Along with emotional-experiential information, Kousta et al. claim that there are two other types of information relevant to the content of concepts. These are sensorimotor information and 'linguistic' information. Sensorimotor information is simply information directly derived from bodily senses. Embodied cognition theorists hold that this information is somehow stored and deployed in cognitive tasks in such a way that its format does not change, and that at least some of the systems devoted to perceiving are also involved in cognition. Kousta et al. (2011) claim that all three types of information (sensorimotor, emotional, linguistic) help constitute the content of both abstract and concrete concepts. The difference between the content of abstract and concrete concepts is a matter of the extent to which each type of information features. Linguistic and emotional information figures more prominently in abstract concepts, whereas sensorimotor information is statistically more preponderant in concrete concepts. Although the nature of sensorimotor information and emotional-affective information is straightforward enough, it is difficult to summarise exactly what Kousta et al. (2011) take 'linguistic' information to be, and this will be the focus of an objection I raise to their account shortly. Now, however, I want to consider a simpler problem with their account of (abstract) conceptual content.

The problem is simply that there are many words with low concreteness scores, but neutral emotional valence scores. Table 2-2 below contains some examples:

Table 2-2 - Emotionally neutral abstract words

Word	Emotional Valence	Concreteness
Imperative	5.21	1.36
Interpretation	5.6	1.4
Concept	4.89	1.41
Extent	5.57	1.44
Theory	5.65	1.47
Motive	5.32	1.5
Intent	5.86	1.52
Fate	5.38	1.53
Paradox	5.4	1.54
Irony	5.19	1.59

Concreteness scores are taken from the Brysbaert et al. (2013) norms, and Emotional Valence sores are taken from Warriner et al. (2013). Warriner et al., like Kousta et al. (2011), use a 1–9 scale for scores of Emotional Valence. A 1 indicates negative emotional affect, 5 indicates neutral emotional affect, and 9 indicates positive emotional affect. Kousta et al. (2011) claim that abstract words tend to be more emotionally valenced than concrete words, and that extreme emotional valence scores are indicative of some kind of emotional, experiential information that is both constitutive of the content of abstract concepts and relevant for the processing of abstract words in lexical decision. Kousta et al. clearly allow that some abstract words are not associated with emotional affect. For example, they hypothesise based on their theory that 'abstract words with affective associations should be acquired earlier than are neutral abstract words' and provide some preliminary statistical evidence in favour of this hypothesis (Kousta et al., 2011, p. 26). An obvious question here is: what constitutes the content of these emotionally neutral abstract words? And relatedly, how could they be acquired?

It seems unlikely that the answer could be sensorimotor information, because one of the defining characteristics of abstract words in their theory is that they tend not to be constituted by sensorimotor information. It is also unclear that emotional affect could explain how these words are acquired, because their defining

characteristic is neutral emotional affect. These items are not associated with any kind of emotionally charged bodily state at all, so it is arguably unlikely that emotional information plays a role in whatever conceptual content they have. It is highly plausible that one of the things that separates the concepts HAPPY and SCARED is that the former is derived from a positive emotional state that comes with ameliorative physiological processes whereas that the latter is derived from a negative emotional state that comes with unpleasant physiological processes; and that information derived or stored from these processes partly constitutes the content of those concepts. However, it is much less plausible that the same thing could be said of the concepts THEORY and IRONY, taken from the table above. A potential response to this point is that even though it may not seem like it, neutral emotional affect really does provide the kind of information required to make this kind of distinction: neutrality of emotional affect provides some kind of content-constitutive information in and of itself. However, I do not think this response will work, because the whole point of Kousta et al.'s (2011) theory is to explain how it is that the content of concrete and abstract concepts differs in such a way as to explain why (emotive) abstract words elicit shorter decision latencies than concrete words. For Kousta et al. (2011), concrete concepts are supposed to come with less emotional affective content than abstract concepts. Degree of emotional affect is operationalised in their experiments as a 1 to 9 scale, with 5 indicating a neutral emotional affect. Abstract items tend to feature at the polar end of this scale, whereas concrete items tend to feature in the middle. If neutral emotional affect is just as content-constitutive as polarised emotional affect, then we have lost both the very measure by which concreteness is supposed to be different from abstractness, and the proposed explanation for Kousta et al.'s experimental results.

It seems like the only possibility we are left with is 'linguistic' information, which brings me to my second objection to Kousta et al.'s account. Unfortunately, it is just not clear what Kousta et al. mean by 'linguistic' information. On the one hand, they note in a section subtitled 'Integrating experiential and linguistic information' that 'we learn a great many words from being told or reading about them' (2011, p. 26). However, they then indicate that they have in mind something like the lexical variables that feature in psycholinguistic experiments. They write:

> …**linguistic** factors such as number of letters, orthographic neighbourhood size, orthographic regularity, and frequency of occurrence also consistently predict lexical decision latencies... Although these variables on their own do not

> account for the abstractness effect… it is important not to discount **linguistic** factors that may relate to processing of abstract and concrete words.
>
> Kousta et al. (2011, p. 26), emphasise mine

In addition, in their closing comments, they suggest that initial concept acquisition by pre-syntactic children is accomplished via world-to-word mapping. A child notices contingences between how the world is at a specific time and what words are spoken, and uses these contingences to learn simple vocabulary like ‘dog’ and ‘mum’. However, once some syntactic knowledge is acquired, children are able to perform structure-to-world mappings that especially aid in the learning of abstract words. These ideas are based on the work of Gleitman and colleagues (Gleitman, 1989; Gleitman et al., 2005). Finally, Vigliocco et al. (2009, p. 223) suggest that linguistic information may also be realised by ‘verbal associations arising through co-occurrence patterns’. Before considering the first three candidates for what counts as ‘linguistic’ information, I want to spend some time on this latter possibility because it is a currently-popular idea in theories of the conceptual system (Barsalou et al., 2008; Connell, 2018; Louwerse, 2011). I am going to argue that ‘distributional’ approaches, as these kinds of view are often called, are not particularly helpful when it comes to explaining the content of abstract concepts.

‘Distributional’ theories of language processing and/or the conceptual system are based on important findings that have emerged in computational linguistics over the last 20 years. In a seminal paper, Landauer and Dumais (1997) demonstrated that with a large enough input text corpus (many millions of words), some well-understood statistical techniques can derive a matrix of ‘distances’ between all of the word types in that corpus. They called their model Latent Semantic Analysis (LSA). This result was important because these output distances more or less tracked human judgements about the ‘semantic similarity’ of the referents of the words. Any English-speaking human can tell you that the meanings of the nouns ‘idea’ and ‘notion’ are similar (or, less controversially, that speakers of English use them to talk about similar things), especially when considered relative to the strength of the similarity that holds between the meanings of the nouns ‘wheelbarrow’ and ‘incantation’. It is remarkable that statistical models can do quite well at this task even when the only information they have comes from examining surface strings in large corpora of natural language, and computing co-occurrence information for pairs of words. The algorithms that are used to solve these tasks have moved on since the

introduction of LSA, and their predictive success has improved. No one can dispute that with a large corpus of text and a cleverly designed statistical model, it is possible to extract a lot of interesting information that at first glance might not have seemed likely to be there. It is also indisputable that the brain is sensitive to some kinds of statistical information when it comes to language processing (see any demonstration of the word frequency effect). However, it is simply not the case that these kinds of model will help us to explain the content of abstract concepts.

The first reason that one might think that distributional approaches can help us to explain anything about the human conceptual system is simply because their predictive accuracy (their ability to track human judgements and succeed at tasks) is high. If a model is successful, the thinking goes, then it must be in some sense isomorphic to the real thing. But this is just a fallacy: the predictive success or aptitude of a model on its own does not provide any evidence that the model somehow 'corresponds to' the phenomenon that is being modelled. At the time of writing, the open-source Stockfish chess engine running on commercially available generic computer hardware is superior in ranking to the current world champion, Magnus Carlsen. Stockfish achieves this by calculating many millions of board positions per second, and selecting the move that optimises some evaluation functions programmed into it by its creators. It does this by representing positions as multiple series of 64 bits, where each type of piece is given its own bitboard, and moves can be computed with bitwise operations on these bitboards. No human player does anything like this. There is no way they could, given that they are human. In fact, historically speaking at least, the better this kind of chess engine got, the less human-like it became, because getting better was often the result of being able to crunch through more board positions per second than before. Human grandmasters calculate relatively few lines for any given position, and they certainly do not analyse anything close to millions of positions per move. Stockfish is excellent at playing chess, but it doesn't provide us a good model of how humans make chess decisions just by virtue of being good at this task.

So what we need is a theoretical argument that shows that distributional approaches do provide a convincing mechanism for explaining abstract conceptual content, and this argument needs to be independent of the predictive success of the models. I will now provide an argument that distributional approaches don't do this. The first thing to note is that not all proponents of distributional approaches believe that they do provide the mechanism we are after. For example, Barsalou et al. (2008, p. 249) state:

> We assume that linguistic strategies are relatively superficial ... Rather than providing deep conceptual information, these strategies provide shallow heuristics that make correct performance easily possible. When the retrieval of linguistic forms and associated statistical information is sufficient for adequate performance, no retrieval of deeper conceptual information is necessary.

For those that do appeal to distributional approaches as a mechanism for abstract conceptual content however, the main problem is that it is just not at all clear how co-occurrence information provides our conceptual system with any content in and of itself. Instead, any content this information provides is just parasitic on other sources that themselves do not explain what abstract conceptual content is. Searle (1980) and Harnad (1990) provide convincing demonstrations of this form of argument, known as the Chinese Room and Symbol Grounding problems, respectively.

In Searle's version, a man who doesn't speak any Chinese is placed in a closed room. He has a large stack of papers next to him with lots of Chinese symbols written on them. He receives pieces of paper from an inbox at one end, on which are written more Chinese symbols, and he has to combine all of his bits and pieces of paper according to some instructions that are written on the wall. The man then places his result in an outbox. Now, it just so happens that his bank of Chinese symbols, the input symbols he receives, and the instructions written on the wall, are adequate to make the output he produces interpretable as a response to the input, for native Chinese speakers. He can do this, even though he's just manipulating Chinese characters according to some instructions written on the wall. Searle thought it was clear that the man does not 'know' Chinese, in any interesting sense: the man couldn't understand or tell anyone what any of the symbols 'meant'. Now, let me extend Searle's thought experiment a bit. Suppose that the man is stuck in the room for a very long time. So long, that he stops needing to refer to the instructions on the wall: he knows the symbol manipulation operations by heart. He also forms associations between the symbols. For example, he comes to realise that symbol A very frequently occurs with symbol B, but that so far it has never occurred with symbol C. He has spent so long in the room that he has a sense of the strength of the pairwise correlations between all of the symbols in his inputs, instructions, lexicon, and output. I contend that the man has not acquired any 'abstract' conceptual abilities whatsoever in virtue of learning these word-form associations. At the risk of belabouring the point: the co-occurrence models that are doing incredible work in computational linguistics

are analogous to the man in the Chinese Room. That does not make them uninteresting or useless. Nothing could be further from the truth. However, it does make it difficult to see why we should believe that distributional approaches are likely to explain how we acquire and think with alleged abstract concepts, such as JUSTICE.

Now, a proponent of a distributional approach to abstract conceptual content may well object to this line of argument. They may object that of course it's not *just* co-occurrence data that constitutes or explains abstract conceptual content. They are allowed to appeal to all kinds of sensorimotor information, plus syntactic information, *plus* the co-occurrence data, in the course of giving this explanation. That's true, but it doesn't seem to me that the co-occurrence data is doing anything in these explanations. For example, shortly I will consider the possibility that syntactic information is what Kousta et al. mean when they talk of linguistic information. We shall see that to the extent that syntactic information *does* provide a mechanism for acquiring abstract conceptual content, it does so completely independently of co-occurrence data. And I do not think an appeal to sensorimotor information will work either: the whole reason for appealing to linguistic information in the first place in order to explain abstract conceptual content was because it is difficult to see how sensorimotor experience could do this.

Now I'll consider the other three candidates here for what linguistic information in Kousta et al.'s (2011) theory should be identified with. It could be whatever information it is that we acquire and store when we encounter linguistic stimuli that we find intelligible. It could be psycholinguistic variable information such as word frequency and orthographic rules. Or it could be syntactic knowledge. Indeed, it could be a combination of all three factors. However, at first glance it seems highly doubtful that the first two candidates can help to provide any account of something like conceptual content, abstract or concrete. To say that the linguistic information that partly constitutes the content of the concept CONCEPT is whatever information it was that we acquired when we heard someone use the word 'concept' seems circular in the extreme. It is precisely the nature of this information that is at issue in the first place. Regarding psycholinguistic variable information, just how could word frequency (or any of the other such variables) be a part of what constitutes the content of the concept CONCEPT, even in principle? It is not clear how tacit knowledge of how many times a word has been encountered (or spelling rules, etc.) could get someone any closer to being able to think using the alleged concepts FATE, or PARADOXES, or INTERPRETATION. The most promising candidate is therefore syntactic information,

and so I will assume that this is the kind of thing Kousta et al. (2011) have in mind when they talk of linguistic information. Referring to the work of Gleitman et al. (2005), Kousta et al. (2011) do briefly sketch out how in principle syntactic information might help children acquire certain concepts after certain 'basic level' concepts have been acquired. This is arguably the most interesting and substantive proposal in their theory, and so I shall devote the remainder of this section to discussing it. I shall conclude that although the ideas that Kousta et al. advance in this respect are plausible in some respects, their theory as it stands just does not seem to apply to a large class of abstract concepts. In part, this is due to a confusing mismatch between what Kousta et al. mean by the term 'abstract' and by what Gleitman et al. (2005) mean by 'abstract', as well as a failure to maintain a distinction between vocabulary acquisition as opposed to concept acquisition. Nothing I shall consider here will falsify their theory: I am simply arguing that it is incomplete.

Gleitman et al. (2005) aim to provide an answer to the question of how children acquire 'hard' vocabulary. As they note, it has been widely demonstrated that children tend to acquire certain types of vocabulary before others. The very first words children learn are universally those words that refer to 'basic level' objects, such as 'dog' and 'mum'. These 'easy' words can be viewed in contrast to 'hard' words, such as cognitive verbs like 'know' and ditransitive verbs like 'give'. These 'hard' words tend to be acquired later. Gleitman et al. propose an explanation of why it is that certain words are hard and acquired late, and why it is that certain words are easy and acquired earlier. I shall give a very short overview of this explanation here, focussing only on those elements that relate to Kousta et al.'s (2011) theory of abstract conceptual content. I will not be able to do full justice to Gleitman et al.'s ideas, which involve a combination of ingeniously constructed experiments and consideration of syntactic theory. In brief, the explanation is that easy word-concept mappings can be acquired simply by observing environmental contingencies in the manner mentioned above. What makes a verb like 'know' harder to acquire than a noun like 'dog' is that environmental contingencies alone simply cannot provide enough data with which to make the required mapping. The claim here is that there are no cues in a child's environment that, on their own, could enable him or her to realise that the English word 'know' should be mapped onto the concept KNOW. However, once some basic level nominal concepts and rudimentary syntactic knowledge is in place, children can integrate information from tactic syntactic knowledge and environmental contingencies, *and* the basic level physical Kind concepts that they have already acquired in order to map more complex vocabulary onto concepts, and in so doing

learn ‘hard’ words. This tacit syntactic knowledge takes the form of constructs like verb subcategorization frames and argument structure. For example, just tacitly knowing that a certain verb takes two arguments and that subjects tend to be causal agents or experiencers provides a surprisingly large amount of information about what concept that verb should be mapped to. Once again, I stress that this is an extremely impoverished formulation of Gleitman et al.’s (2005) account. However, we are now in a position to assess the extent to which Kousta et al. (2011) are can rely on these ideas to relate ‘linguistic’ information to the content of emotionally neutral abstract concepts.

The first thing to note is that, although there is some overlap, Gleitman et al. (2005) mean different things by the terms ‘abstract’ and ‘concrete’ than Kousta et al. do. Gleitman et al. (2005, p. 26) make this explicit in a footnote. They point out that although it is widely accepted that the term ‘concreteness’ is associated with vocabulary referring to basic level entities and that these basic level entities have some properties that makes their accompanying vocabulary easier to learn, it is also the case that there are very many concrete entities that are more difficult to learn vocabulary for than others. For example, partitives (‘trunk’ versus ‘elephant’), superordinates (‘animal’ versus ‘dog’), proper names (not realising that ‘Daddy’ is a name for a specific man), and ‘situation-restricted’ terms (‘passenger’ versus ‘man’) are all hard vocabulary in that they are acquired relatively late (examples taken from Gleitman et al. (2005)). They state that ‘concreteness is itself a term in need of considerable sharpening’ and that they are simply using this term as a ‘nickname’ for whatever properties there are that ‘easy’ vocabulary turns out to have (Gleitman et al., 2005, p. 26). So for Gleitman et al., while concreteness does certainly have some connection with some physical essence of items in our conceptual repertoires, they really want to reserve the term for ‘easy’ vocabulary. By the same token, in their narrative, Abstractness is just whatever properties there are that make some words ‘harder’ than others to acquire.

This terminological nuance makes it difficult to relate some aspects of Kousta et al.’s theory to that of Gleitman et al., because it does not seem like they are using concreteness and Abstractness to refer to the same properties. And furthermore, while Kousta et al. are trying to explicate what it is that constitutes Abstractness (i.e. emotional affect), Gleitman et al. seem to not want to commit to anything on this front. Indeed, for most of their article, they prefer to use the terms ‘easy’ and ‘hard’ instead of ‘concrete’ and ‘abstract’. This mismatch between different notions of Abstractness can be seen in Kousta et al.’s (2011, p. 26) discussion of the early acquisition of

certain abstract words referring to emotional states. They note that the words ‘good’ and ‘happy’ tend to be acquired at a relatively young age. But it is precisely the fact that these words are acquired early that makes them *not* abstract in the sense favoured by Gleitman et al. (2005). However, despite this confusion, Kousta et al.’s suggestion that basic level emotion words such as these are acquired on the same basis as basic level Kind words such as ‘dog’ is interesting and plausible. They suggest that ‘abstract words denoting emotional states, moods, or feelings also fall in the same category [as basic level physical Kind words] for which a mapping from the word to the world (albeit the internal world) is possible’ (2011, p. 26). So although their claims may carry some weight with respect to some abstract concepts, we still do not have an explanation for how emotionally neutral abstract concepts are acquired and constituted. That Gleitman et al. (2005) do not provide this explanation, and did not intend to in the first place, can be seen by considering two more points of confusion regarding the terminology that features in the theories under discussion. Firstly, Gleitman et al. almost exclusively focus on the acquisition of certain verbs with complicated argument structures and subcategorization frames, such as ‘know’ and ‘chase’, and their aim is to explain the acquisition of these verbs with reference to syntactic theory in such a way as to account for what makes them ‘hard’ to learn as opposed to ‘easy’. The emotionally neutral concepts that I take to be problematic for Kousta et al.’s (2011) account are simple nominal concepts such as IDEA or MOTIVE. It could be the case that an explanation regarding the content and acquisition of these concepts can be made in terms of tacit syntactic proficiency in recognising argument structure and subcategorization frames, but we do not have such an explanation yet.

The second point of confusion here is between concept acquisition and vocabulary acquisition. I take Kousta et al. (2011) to be offering an explanation of the content and acquisition of concepts. Roughly, the question might be phrased as, ‘how does a child come to understand what anger is, and/or think using the concept ANGER?’. But this is a separate issue to the question of how it is that a child acquires the *word* ‘anger’ and its syntactic properties. Gleitman et al. (2005, pp. 25–26) make the explicit claim that ‘a considerable part of the bottleneck for vocabulary learners is not so much in limitations of the early conceptual repertoire but rather … determining which phonetic formative expresses which concept’. Gleitman et al. assume that a child must have already acquired a concept in order to be able to map it to the appropriate vocabulary, and cite a range of studies that purport to show that infants have acquired concepts before they have acquired the relevant vocabulary. Indeed, they specifically take aim at theories that assume that concept acquisition just is

vocabulary acquisition. Their goal is to provide an explanation of how concepts that a child has already acquired can be mapped onto vocabulary that they do not yet know. Recall that Kousta et al. hold that 'linguistic' information partly constitutes the content of abstract concepts. I do not want to argue that this claim is definitely false, but rather I want to point out that they have not provided an explanation of what this linguistic information is or how it does provide concept constitutive information. The problem is that it is not obvious how linguistic 'syntactic' information could help constitute the content of concepts that *need to have been already acquired* in order to acquire this syntactic information. As far as I can tell, Gleitman et al. (2005) do not provide this explanation.

An important question we have not considered so far is: *why* would the emotional affective information that is supposedly constitutive of an abstract concept speed up lexical decision time? This is similar to the question we asked previously about explanations for why concrete words supposedly elicited shorter reaction times than abstract items. What is the process by which the positive emotional experiential information constitutive of the concept HAPPY speeds up a lexical decision to the letter string <happy>? If there is such a process, then this would imply something quite remarkable about how the human brain stores information relating to orthographic word forms, namely that the availability of a word form to the processing and decision procedures of the mind-brain is ranked in order of degree of emotional valence of its associated concept. It could be that Kousta et al. (2011) have such an explanation, or independent evidence that this implication is true, but they do not seem to provide one in their text. In a reply to Paivio's (2013) commentary, Vigliocco et al. (2013) do briefly consider this issue, but I am not sure that their explanation suffices. They state:

> … one can argue that lexical decision is not an appropriate task to assess predictions by DCT because lexical decision only taps into early processes, whereas the engagement of the imagistic system would occur later (Paivio, 2013, pp. 282). Semantic effects are, however, consistently observed in lexical decision, whether reflected in the thousands of studies in which semantic priming is observed… or reflected in numerous other studies showing the effects of semantic variables at the single word level.
>
> Vigliocco et al. (2013, p. 289)

It is true that ‘semantic’ effects have been frequently reported in lexical decision studies, but it is important to note that there is some debate as to what produces these effects. For example, Shelton and Martin (1992) argue that priming in lexical decision does not really reflect a ‘semantic’ effect at all, but is due instead to distributional relations between lexical items (the word ‘cat’ primes ‘dog’ because ‘cats’ is frequently followed by the words ‘… and dogs’, but not because of some overlap in content between the concepts CAT and DOG). Vigliocco et al. (2013) cite Lucas’ (2000) review on semantic priming in support of their case, but even if we accept that such priming truly is ‘semantic’ in the sense that it results from a causal effect of some conceptual content on the availability of other conceptual content, it is still not the case that this could explain Kousta et al.’s (2011) results. This is because Kousta et al. (2011) simply did not run any experiments in which priming should arise at all. They presented their stimuli in a randomised order for every participant, and any prime-target relations that occurred would have been accidental. Indeed, one of the reasons that the presentation of stimuli in lexical decision tasks is randomised is to eliminate the effects of any chance ordering of word pairs that might give rise to priming that is not of experimental interest.

Vigliocco et al. (2013) also cite Balota et al. (2007), Chumbley and Balota (1984), and James (1975) as examples of different kinds of ‘semantic’ effects in lexical decision, which are not based on priming. However, I just can’t see how these particular studies provide support to the idea that high emotional valence speeds up lexical decision latencies. Chumbley and Balota (1984) found that there was a statistically significant correlation between the time it takes participants to produce an associated word, given a stimulus word, and lexical decision latencies to those stimulus words. This is a completely different kind of ‘semantic’ effect to the one that Vigliocco et al. (2013) are positing (that the availability of a word form to the processing and decision procedures of the mind-brain is ranked in order of degree of emotional valence of its associated concept). Balota et al. (2007) is a report of the English Lexicon Project, which is a database of lexical decision reaction times. But none of the other variables included in that project seem to be ‘semantic’; instead they are patently non-semantic variables such as word length, frequency, and orthographic neighbourhood density. Finally, James (1975), as we saw above, investigated concreteness effects on lexical decision, and did *not* consistently find such effects.

The upshot of all of this is that even if we accept that abstract words *do* elicit shorter decision latencies than concrete words, we have not yet achieved a satisfying

explanation of why this difference should occur. Furthermore, as we saw above, there are other lexical decision studies that purport to show a decision latency advantage for concrete words (Binder et al., 2005), as well as studies that suggest that there is no difference (Brysbaert et al., 2016). I take all this to show that, for the present moment at least, we just should not rely on lexical decision data in order to inform our theories of concepts with respect to the concrete-abstract distinction.

2.9 Emotion concepts and the concrete-abstract distinction

I will now offer a relatively self-contained argument against the validity of the concrete-abstract distinction as it applies to emotion concepts. I appreciate that I have been quite critical of Kousta et al. (2011) so far. But I want to stress that I think their proposals regarding emotion concepts such as ANGER, FEAR, and HAPPY are interesting and plausible suggestions. As mentioned above, accounts of abstract conceptual content tend to be thin on the ground, and so any progress in this regard is something to welcome. However, I think that if we examine the logic of the concrete-abstract distinction with respect to Kousta et al.'s ideas, the motivation for making the distinction in the first place collapses. As we saw in the introductory sections of this chapter, abstractness is rarely defined in a rigorous way. Our concrete concepts have something in common, which is that they are presumably acquired on the basis of sensorimotor experience. However, the only property that abstract concepts seem to have in common is that they 'aren't concrete'. I now want to argue that, actually, we have no reason to hold that emotion concepts aren't concrete. In fact, Kousta et al. have provided a clear account of exactly how it is that emotion concepts *aren't abstract.*

Many theorists (e.g. Paivio (1986), Dove (2016, 2011) Barsalou (2003), and Kousta et al. (2011)) emphasise that sensorimotor experience is a central source of conceptual content. However, all of these theorists, arguably even Paivio, are also more or less explicit that sensorimotor experience should not be thought of as relating only to the five canonical bodily senses. Although Kousta et al. stress the point the most, Barsalou, Dove and Paivio all make mention of introspective and 'affective' states too. Barsalou, and Kousta et al. both explicitly claim that the mind-brain can 'capture' information contained in affective mental states. The key element of Kousta et al.'s theory of abstract conceptual content is that the mental representations 'captured from' these affective mental states help to constitute the content of abstract concepts, and not concrete concepts. Barsalou (1999) suggests something similar.

In this section, I argue that this move is a mistake. If this idea about capturing information derived from affective states is even close to correct, then the way that the concrete-abstract distinction tends to be drawn is not only vague, but it is inconsistent. And if a distinction is vague and inconsistent, then we should probably abandon it. My view is that it is not consistent for these theorists to maintain that DOG is concrete in a way that emotion concepts are not. I shall argue for this view by providing a rough characterisation of how embodied cognition theory says we acquire the concept DOG in terms of sensorimotor experience, and then showing how the same characterisation could just as well apply to FEAR (an emotion concept). I then consider two potential objections, and some responses to them.

Imagine a human, Isobelle, who has never had any experience with any dogs or dog-like animal, but has otherwise typical cognitive function. One day, Isobelle encounters a dog. Light reflected off of the dog travels through the lens of Isobelle's eye and hits her retina. There, a formation of light sensitive cells passes this sensory impression along the optic nerve in the form of action potentials and neurotransmitters released across synaptic gaps, towards the visual cortex in the occipital lobe. The dog barks. This barking causes compressions in the air in the form of a wave. This wave strikes Isobelle's ear drum, and the resulting vibrations are propagated through a complicated series of tiny bones until they reach the cochlea. Spirals of miniscule hairs arranged by their responsivity to different frequencies then pass this sensory information to the auditory nerve, towards auditory cortex in manner analogous to the light sensitive cells of the retina. The information instantiated in various neural assemblies in the visual and auditory cortex that were stimulated as a result of the physical process of experiencing the dog starts to become 'bound' together as a concept. This is how Isobelle acquires the DOG concept. When Isobelle thinks about dogs, this neutrally instantiated information is what allows her to do so, by virtue of the fact that it is *sensorimotor* information derived from her physical experience(s) of encountering a dog. The fact that Isobelle's DOG concept was acquired this way, through a physically instantiated interplay between an external stimulus, cells, and electrochemical signals, means that neuropsychologists label it a 'concrete' concept.

Barsalou, Dove, and Kousta et al. are not exactly clear on how, or at what point in these strings of events, the sensory information is 'captured' and made available to cognitive processes. But if the embodied cognition account is true then hopefully at some point in the future, the fields of neuroscience and neuroanatomy might be able to tell us. Now, let's assume this account is indeed broadly correct. I

think that there is no basis for claiming that FEAR is any less 'concrete' than DOG. In order to demonstrate this, let's imagine that Isobelle has never been afraid before:

One day, Isobelle is suddenly and without warning plunged into total darkness, and hears an ear-splitting animal scream accompanied by ominous industrial noise. In response to this unexpected stimulus, Isobelle's sympathetic nervous system and hypothalamus release a cocktail of chemicals and hormones through her body and bloodstream, such as adrenaline, noradrenaline, and dozens of others. These chemicals cause, among other things, her heart rate to quicken and her blood pressure to rise. Her breathing also speeds up. Various muscles tense up, and other muscles relax. Her blood sugar level rises. This sudden combination of muscle tensing and chemical bombardment causes a hot, nauseous feeling that seems to emanate from her stomach. Veins are constricted so that blood is diverted away from her extremities and towards major muscle groups, causing a characteristic chilling of her fingers and toes. All of these sensations, among others, are ultimately transferred through nerve cells and received as electrochemical signals in her brain distributed through assemblies of neurons. It seems plausible to me that this is at least part of the story of how Isobelle comes to be able to think about being afraid: she has experienced fear. The information instantiated in neural assemblies that were stimulated in response to fear-inducing events can capture this experience and become bound together as a concept. When Isobelle thinks about being afraid, this neutrally instantiated information is what allows her to do so, by virtue of the fact that it is *sensorimotor* information derived from her physical experience of fear-inducing events. My argument is that the fact that Isobelle's FEAR concept was acquired this way, through a physically instantiated interplay between an external stimulus, cells, and electrochemical signals, means that we have no grounds for wanting to claim that FEAR isn't concrete.

Suppose that the sketches just provided are a reasonable (but partial) summary of how embodied cognition accounts explain concept acquisition. I do not see much reason to suppose that the neural assemblies that responded as a result of the chemical and physical activity produced by the fear-inducing event 'capture' this experience in a different way to how the neural assemblies in visual cortex can capture the experience of light reflecting off of a dog. Kousta et al. (2011) consistently indicate that emotion concepts are 'abstract' and therefore qualitatively different to 'concrete' concepts. But it seems to me that this is the wrong way to go. Much of the discussion about the qualitative distinction between concrete and abstract concepts concerns how they are 'represented in' the brain. The general consensus is that they

must be ‘represented differently’. Often, the kind of evidence that gets deployed in support of this consensus is based on measurements of neural activity (Barber et al., 2013; Binder et al., 2005; Holcomb et al., 1999; Kounios and Holcomb, 1994). But I think that the FEAR story shows that any of the concepts that theorists refer to when they render emotion words in capital letters (FEAR, ANGER, LOVE, etc.) just should not be thought of as abstract at all. That is because, from the point of view of electrochemical signals propagating throughout assemblies of neurons and nerve cells, there isn’t any reason to suppose that the content of thoughts that feature DOG is ‘represented differently’ to the content of thoughts that feature FEAR (or any other affective concepts).

I will now consider some objections to the point I have been trying to make. Perhaps the most obvious objection is that there is still a clear difference in how FEAR and DOG are mentally represented: neuroimaging investigations show that spatially distinct structures in the brain respond to reading emotionally laden words such as ‘fear’ and (more) emotionally neutral words such as ‘dog’ (Maddock et al., 2003). And if different neural structures are implicated, then surely that indicates that there are qualitatively different kinds or formats of representation at play, because of the spatial distinctiveness principle we saw before (Holcomb et al., 1999). So perhaps we might maintain that FEAR is importantly distinct from DOG on these grounds, and we might as well label this distinction using the familiar concrete-abstract terminology. However, I don’t think this objection goes through if we want to maintain the rest of the claims made by embodied cognition theories. That’s because there are spatially distinct structures implicated in processing all kinds of stimuli, and there doesn’t seem to be any principle that separates the structures implicated in experiencing emotions (or reading emotionally-laden words) from the structures implicated in perceiving dogs, other than sheer stipulation.

Consider again the story about Isobelle and her DOG concept. The auditory and visual cortex are pretty spatially distinct. The visual cortex is located towards the back of the brain, in the occipital lobe. The auditory cortex is located on the upper side of a portion of the temporal lobe, in the lateral part of the brain. Neuroscientists can reliably localise neural responses to visual and auditory stimuli to these regions. My point here is that the same argument about spatially distinct structures representing the DOG and FEAR concepts should just as well apply to the different kinds of sense experience we have of dogs. If spatially distinct structures indicate different kinds or formats of representation (or different kinds of concept), then we should say that there is a DOG(AUDITORY) concept, and a DOG(VISUAL) concept,

and a DOG(TACTILE) concept, and so on, as opposed to a single DOG concept. But this idea does not seem to be popular in the literature. Rather, the prevailing assumption is that there is just one DOG concept, and that it is made up of both auditory and visual mental representations, among others. Again, I think it would just be stipulative to claim that visual and auditory mental representations are 'concrete' but affective mental representations are qualitatively different and 'abstract', when the acquisition of the concepts they constitute can plausibly be traced back to the same kinds of biophysical processes. Moreover, there are other pairs of domains for which different neuroanatomical regions are implicated in processing or categorisation of stimuli. Neuroscientists find that different areas of the brain preferentially respond to seeing pictures of animals versus pictures of tools (West et al., 2001). But, as far as I know, no one has been tempted to suggest that this indicates that there is a fundamental distinction between the formats of the mental representations that support the visual categorisation of tools and those that support the visual categorisation of animals.

A counterargument here could be that there are some bodily sense apparatuses that are primary in some respect. Perhaps the five canonical bodily senses; sight, sound, taste, touch, and smell. So it could be that a concept is concrete if it is acquired and mentally represented on the basis of some primary sense data coming from these five senses. The FEAR concept wouldn't be counted as concrete because it was acquired on the basis of some non-primary, body-internal sense data (such as the feeling of hot nausea caused by the contraction of various muscles and the diversion of resources away from the digestive system during bouts of fear). I have two responses to this counterargument. Firstly, I think it is still purely stipulative. We would need some evidence or reason to believe that stimulation of body-internal nerve cells produces qualitatively different kinds of mental representation to stimulation of body-external nerve cells. Secondly, without this evidence, this stipulation doesn't take on any explanatory role in maintaining the concrete-abstract distinction. The only theoretical element that this stipulation preserves is the concrete-abstract distinction with respect to DOG and FEAR. And the very point I am arguing is that in this case, there might not be a need to draw the distinction in the first place.

There is another issue that I want to consider here, and it has to do with the often-proposed relationship between 'linguistic' representations and abstract concepts. Paivio (1986) proposed a kind of linguistic representation, the logogen, in an attempt to explain the content of abstract concepts. Dove (2011, 2014) especially seems convinced that 'linguistic' representations are the key factor in explaining what

abstract conceptual content is and how it is acquired (solving the 'problem of abstraction'). As we saw above, Kousta et al. (2011) propose that 'a statistical preponderance of affective and linguistic information [underlies] abstract word meanings' relative to concrete word meanings. Let us allow for a moment that 'abstract' conceptual content does distinctively and importantly depend on some kind of linguistic representation, and that this is partly what separates abstract conceptual content from concrete conceptual content. Now, there is even *less* reason to suppose that FEAR is abstract in a way that DOG is not (or, conversely, that DOG is concrete in a way that FEAR is not). This is because the FEAR story shows that Isobelle's experience of fear, and subsequent simulation of it, doesn't have to have anything to do with linguistic processing at all. It seems implausible to suppose that we can't have a concept of a given emotional or affective state until we have learned that there is a word in our language that refers to that state. So there is nothing for a 'linguistic' representation to do when it comes to acquiring a concept of an emotion, or what it is to be in a certain emotional state. The upshot of all of this is that for any affective state for which we want to posit a concept, that concept should not be labelled abstract: there is no principled ontological distinction in embodied cognition between concepts of dogs and fear. And, I want to stress that this is actually a positive outcome, because if you endorse an embodied cognition account, the set of concepts that has traditionally been considered hard to explain is now smaller.

2.10 Summary of Chapter 2

I now summarise what we have seen so far. Concreteness is an important psycholinguistic construct that has been investigated extensively for decades. In psycholinguistic experiments, concreteness is a property of words, and words are assumed to have close and reliable connections with the concepts we have. A concrete concept is a concept that has been acquired on the basis of sensorimotor experience. What an abstract concept is has proven very difficult to define, although theorists tend to associate them with 'linguistic' and 'affective' mental representations. In list memory experiments, concreteness effects are reliably obtained whereby concrete words are easier to remember than abstract words. In EEG experiments, concrete words reliably elicit larger N400 amplitudes than abstract words, despite the fact that N400 amplitudes are thought to correlate with processing difficulty, and the typical assumption is that concrete words are easier to process than abstract words. Although statistically significant experimental contrasts are often obtained, the fMRI data on concreteness effects is highly variable and there are many inconsistent findings. In lexical decision experiments, results are even more variable. Sometimes

these experiments produce an advantage for concrete items, sometimes they produce an advantage for abstract items, and sometimes there is no concreteness effect at all. This arguably invalidates theoretical explanations predicated on the idea that concrete words elicit shorter reaction times than abstract words in these tasks (Binder et al., 2005; Connell and Lynott, 2012). Kousta et al. (2011) obtained evidence for an advantage for abstract items, and proposed that this advantage is due to the fact that abstract concepts are constituted by information derived from emotional affective states to a greater degree than concrete items. Even if we were to favour their finding over the dozen or so other conflicting findings in the literature, I have suggested that, without some explanation of what the causal effect of emotional valence is on decision latencies, we should be cautious about endorsing their theory as it stands. I have also presented an argument that if we accept Kousta et al.'s (2011) fundamental point that information derived from emotional affective states can provide the mechanisms for acquiring concepts and instantiating them in cognitive processes, we should not hold that these particular concepts are abstract. That is because these concepts seem to have the same properties as concrete concepts, and so with respect to emotion concepts, there is no reason to draw the concrete-abstract distinction.

In the next chapter, I consider how theories of cognition and concepts intersect with the concrete-abstract distinction. I will argue that no matter which theory we endorse, some proposed abstract concepts do not actually do any explanatory work in these theories. Consequently, we should not hold that words and concepts stand in a reliable correspondence with one another. Furthermore, this correspondence is least reliable when it comes to those words which we tend to label as 'abstract'. I think this argument provides evidence against the utility of the concrete-abstract distinction, as it is operationalised in psycholinguistics.

Chapter 3: Concreteness and concepts

In the previous chapter, we considered concreteness as a psycholinguistic variable; the experimental effects that have been attributed to it; and some explanations of these experimental effects. We also saw that pretty much universally, psycholinguists assume that words of English match up with concepts in a regular and relatively automatic way, such that when a participant encounters a word in an experimental task, they will undertake processing that operates over a *concept* that corresponds to that word. In this chapter, I consider various theories about what concepts are, what properties they have, and what role they play in cognition. I show that philosophers and theoretically-minded psychologists tend to agree with psycholinguists about the relationship between words and concepts. I will pay particular attention to two kinds of view: a Fodorian (1998, 1975) language of thought and a Barsalou-ian (1999; 2008) simulator account. I single out these kinds of view because in an important respect they represent opposing schools of thought about the properties that conceptual mental representations have. Fodor believed that concepts are amodal, arbitrary symbols that enter into syntactic relations with each other in order to produce thoughts. Barsalou believes that concepts are *simulators* of experience, and that the mental representations involved are not arbitrary, in the sense that they are the same or similar in format to the mental representations involved in the experience from which a simulator was derived. I want to convince you that no matter which kind of view you want to endorse, the concrete-abstract distinction does not actually amount to anything. I will consider the case of the alleged concept JUSTICE, an abstract concept par excellence. I will demonstrate that it is explanatorily vacuous both as a symbol in a Fodorian language of thought, and as a simulator in a Barsalou-ian embodied cognition framework. By this, I mean that as a component of either theory, JUSTICE does not allow us to explain or predict anything about human behaviour or cognition. Since this is the main job of a concept, I conclude that there is no such concept, and that therefore the relationship between words and concepts is not as reliable as it is generally assumed to be. I end the chapter by acknowledging two major objections to the arguments I am about to make. The chapters following this one contain my responses to these objections.

The structure of this chapter is as follows. To start, I provide a summary of Fodorian and Barsalou-ian views on concepts and cognition. I also briefly consider a range of other views. I show that although there is certainly disagreement when it comes down to the details, there is actually quite a strong consensus across the philosophical and psychological literatures regarding three non-trivial issues. With this general understanding of the differences and similarities between Fodor and Barsalou in place, I move on to make my arguments against the explanatory utility of JUSTICE. I show that every theory we have considered so far has a reasonable story to tell about concepts we think of as concrete, such as DOG. Theories can explain how DOG is acquired, and they have sensible proposals about what properties DOG has, such that it can play certain roles in cognition and behaviour. However, there is no theory that has a reasonable story to tell about the alleged concept JUSTICE, and there are no sensible proposals about how JUSTICE is acquired or what properties it has such that it plays a role in anything. I suggest that we simply have not been thinking about concepts in the right way. Concepts are parts of theories of cognition: they are theoretical posits. We use concepts to explain behaviour and cognitive processes. I rehearse some thought experiments designed to show that the reason that we cannot say how JUSTICE is acquired, or what properties it has, is that JUSTICE does not play an explanatory role in our cognitive theories. If a theoretical posit does not play an explanatory role then it does not belong in a theory, and we should abandon it. I conclude that we should not posit JUSTICE, and I suggest that the same strategy I employ here might be used to rule out other alleged abstract concepts which cause trouble for our theories of cognition.

3.1 The Fodorian Language of Thought Hypothesis

Typically, when psychologists talk about concepts, they (knowingly or not) draw on a framework inspired by Fodor (1998, 1975). They draw on this framework in two ways. Firstly, Fodor took word-level linguistic structures to reveal concepts that we possess.[2] So one way of deciding what basic-level concepts there are is to see if there is a word for something in the language(s) that we speak. On this approach, it's

[2] The reader (or indeed Fodor himself) might object to this characterisation of his view. It might be difficult to pin Fodor down on what he thinks the precise relationship is between the number of words we have and the number of concepts we have. I would point out that others have also suggested that this is the most straightforward way to read him (Margolis and Laurence, 2015; Rescorla, 2017; Sperber and Wilson, 1998).

natural to suppose that English speakers have a JUSTICE concept because there is an English word, ‘justice’. Spanish speakers also have a JUSTICE concept because there is a Spanish word, ‘justicia’. Note how the distinction between words and concepts is supposed to work: both English and Spanish speakers have a JUSTICE concept, and not different concepts, despite the fact that each language has a different word for justice. This property of concepts is called ‘publicity’: concepts are ‘public’ in the sense that two people can (and generally do) have the same concept Type in their respective conceptual repertoires. The view that publicity is a necessary property for concepts to have is endorsed fairly frequently (Fodor, 1998; Prinz, 2004), although Barsalou (1999, 2017) is harder to pin down on how public he thinks concepts are: he sometimes says that ‘simulators’ - his term for concepts - are ‘shared’ but other times he uses weaker phrases such as ‘highly similar’. Publicity is frequently endorsed because it’s often supposed that concepts must be shared in order to explain how it is that humans understand and reason about each other. If we didn’t have the same concepts, the thinking goes, then we couldn’t communicate properly, because if we didn’t have the same concepts, then we would be incapable of thinking the same thoughts. I return to this issue later on, but for now, it’s fair to say that the word-concept approach is intuitively appealing because it suggests why some concepts seem ‘basic’ in some sense; it suggests a way of delineating a kind of core conceptual library. No one has a basic-level SUSPICIOUS UNCLE LIVING IN OHIO concept because there isn’t a lexemic structure in any language (I assume) that corresponds to whatever it is that SUSPICIOUS UNCLE LIVING IN OHIO applies to. As we will see, pretty much every theory of concepts assumes that words have a close connection to concepts.

The second main contribution Fodor made to psycholinguistic theorising about concepts is his notion of what the relationship between thoughts and concepts is. Fodor was very keen to stress that concepts are ‘compositional’. This means that they are ‘constituents’ of thoughts, and they combine in systematic ways to produce the content of the thoughts of which they are a part. So thoughts are made up of different concepts. In turn, these constituent concepts may themselves be composed of concepts. Ultimately though, thoughts will decompose down into atomistic concepts that are not themselves made up of constituent concepts: these are the basic-level concepts that a natural language might more or less reveal. Fodor viewed concepts as atomistic in a relatively strict sense: on Fodor’s view, there is no level of analysis below that of the concept. For him, the concept is the fundamental unit of cognition, and by ‘fundamental’ I do not simply mean ‘most important’; I mean that basic-level

concepts are the terminal of the hierarchy of cognitive constructs. Modern psychological approaches to concepts use the idea that words reveal concepts, and that concepts compose, but in general they do not treat concepts as atomic. Indeed, much of the current work on the concrete-abstract distinction is based on the assumption that concepts must be 'made up of' other mental representations (e.g. sensorimotor representations), and one of the key challenges is to specify what kinds of mental representations make up each of our concepts, and how they do so.

In any case, on Fodor's view, concepts are units in the language of thought. Concepts can enter into *syntactic* relations with one another in order to produce thoughts. If the concepts DOG, LOVE, and others enter into the right syntactic relations with one another, specified in the language of thought, then a cognitive agent has a thought expressible with the sentence, "I love dogs". This syntactic relationship between concepts is a crucial component of Fodor's view because he claimed that this is the only way of accounting for what he called the 'productivity' and 'systematicity' of thought. Thought is 'productive' in that we seem to be able to entertain an unlimited number of them. Thought is 'systematic' in that that are important regularities about the thoughts we are able to have. If someone can think a thought expressible with the sentence "I love dogs", then, Fodor suggested, they will *always* be able to think a thought expressible with the sentence "dogs love me". Fodor believed that the only way for these two properties to obtain was if the structure of thought was relevantly similar to the structure of language; which is to say that thought has a syntactic structure, of which concepts are constituents.

Very importantly, on Fodor's view concepts are really just *arbitrary symbols.* We use English words ("dog") to depict them, but the concepts that feature in cognitive processes do not have any content in and of themselves. There is no necessary relationship between the sensorimotor experience from which a concept was acquired, and the format of that concept as a mental representation: this makes Fodor's theory an *amodal* view of concepts. You might wonder how this could possibly be the case. Surely a concept must have some kind of content, because concepts must apply to things out there in the world. If a concept is just an arbitrary symbol, not completely unlike a series of 1s and 0s in machine language, then how would a concept do this? Another important component of Fodor's theory, which we shall return to later, is the mechanism by which concepts are 'individuated'. To say what individuates a concept is to say what makes it the case that a concept *is* the concept that it is. In other words, what makes it the case that a DOG concept *is* a concept that picks out dogs, rather than, say, cats, or theorems? Fodor's answer to this tricky

problem was that the thing that individuates the DOG concept is that it is *nomologically locked* to dogs, out there in the world. There is a law-like connection between DOG and dogs. The DOG concept, an arbitrary symbol, is tokened by dogs; it is causally related to dogs in a very specific way. That's what makes a DOG concept a DOG concept, according to Fodor. This solution to the problem of concept individuation has a potentially surprising upshot. You might ask how or why the human mind-brain becomes nomologically locked to something, or anything, in the first place. Why do arbitrary symbols in the language of thought become nomologically locked to certain categories and not others? Fodor's (1975) answer to *this* problem was to assume that many concepts, and/or perhaps the nomological lockings themselves, are innate. Our mind-brain just *is* the kind of thing that becomes nomologically locked to the categories that it does in fact become locked to.

3.2 Barsalou's simulator theory

Barsalou's (1999; 2008) account looks quite different to Fodor's. In brief, Barsalou's view is that neural assemblies that respond to perceptual stimulation are able to store aspects of the experience of this stimulation for later use. Repeated exposure to categories of objects and events results in a distributed network of stored sensorimotor representations that can be called upon to simulate the experiences from which they were originally derived. So, for example, after I have experienced a dog in various ways (seeing one, hearing one, etc.), a distributed assembly of neurons captures this experience and becomes bound together in a simulator. On this view, having a DOG concept is having a simulator — a distributed neural assembly containing sensorimotor representations of experience of dogs — that is capable of partially simulating this experience. It is these modal, sensorily-bound simulations that constitute (or generate) the units of thought.

There are number of important caveats and nuances in Barsalou's theory. In the first instance, Barsalou (1999) is careful to stress that a simulation will always be imperfect in various ways. So, a specific simulation of dogs that features in a thought I have *about* dogs will always be a somewhat vague approximation of the experience from which I constructed the dog simulator. This imperfection results from the fact that simulators and simulations are 'dynamic' in that they are constantly changing. Any new experience of dogs has the potential to change the structure of my existing dog simulator. Furthermore, every specific simulation of a dog that my dog simulator produces is highly task- and context-dependent. If simulations from different simulators interact with each other in a certain way, then it is even possible for these

simulations to change the structure of the simulators that generated them via a feedback mechanism. On this view, every time I have a thought that is about dogs, the specific simulation of dogs that is instantiated is temporary and 'one-off'. In the general case, the simulations that feature in thought A will not be the same as the simulations that feature in thought B, even if these simulations come from the same simulator(s). This is because when a simulator is active (generating a simulation that will feature in a cognitive process), only a subset of the total network that constitutes that simulator will ever be engaged, and this subset is so task- and context-dependent that it will never be perfectly re-instantiated. Typically, networks of simulators will be activated during cognitive processing in such a way that there may be an enormously complicated interplay between parts of different simulators, feedback between simulations and simulators, dynamically changing simulations, and so on.

Another important point to stress is that, as Barsalou (2016) has recently emphasised, he does not use the term "sensorimotor experience" in a narrow way. He has never claimed that experience in the five canonical bodily senses is the only thing that can play a role in the acquisition of concepts, or their instantiation in thoughts. Instead, he allows that the whole range of information derivable from any bodily state can constitute simulators that play these roles (such as emotional-affective states, interoceptive states, and proprioceptive states). Probably the most important difference between Fodor and Barsalou is that Barsalou endorses an *embodied,* or *situated*, cognition view. On this sort of view, the mental representations that feature in thoughts are not amodal: they are relevantly like the mental representations produced by actual sensorimotor experience of an object or event. They retain this sensorimotor character when they feature in cognitive processes that take place in the absence of these objects and events. For embodied cognition views, the thing that makes a DOG concept a DOG concept is that it is made up of sensorimotor representations derived from perceptual experience of dogs. So there is a *non-arbitrary* connection between the format of the mental representation, DOG, and the category it applies to, namely dogs.

3.3 Areas of agreement about concepts

We will return to specific details of both of these views when it comes to my arguments against the explanatory utility of JUSTICE, but for now I just hope to have outlined in broad strokes what each kind of theory involves. Drawing on a range of other thinkers and proposals, I now want to show that in some respects, there is actually quite widespread agreement about what properties concepts and thoughts have in both the

psychological and philosophical literatures. In what follows, the term 'unitary cognitive resource' will be crucial to much of my exposition, and so I introduce this section by outlining what I mean by it. By 'unitary cognitive resource', I mean a set of mental representations, and the brain structures that underwrite them, that form a cohesive whole *in the context of a psychological theory*. The stress in the previous sentence has intended significance: many views of concepts might hold that as neuro-psychologically instantiated objects, concepts aren't unitary or cohesive. However, I am referring here to concepts as elements of our psychological theories and explanations. I appreciate that this distinction might seem somewhat obscure, so let me try and flesh it out a bit:

Unitary cognitive resources can occur in thoughts, and they have properties that allow them to play causal roles in the thoughts in which they occur. If a philosopher or psychologist interested in concepts has rendered a word in capitals, italics, or quotation marks, then it seems to me quite likely that they are alluding to what I mean by a 'unitary cognitive resource'. The idea is that my DOG concept is a unitary cognitive resource in that whatever structures and processes there are in my brain that enable me to reliably categorise and cognise about dogs as such, they are sufficiently cohesive across different instances of my categorising and cognising about dogs that we can try and build a theory of them. With sufficient knowledge of the brain and the appropriate technology, perhaps we could in principle (although maybe not in practice) specify exactly which structures in my brain are the ones that underwrite my ability to categorise and cognise about dogs as such, or we could construct an algorithm that specifies how these structures will change in response to stimuli and other processes taking place within my brain. There is some kind of 'core' to these structures, and the mental representations that they underwrite, that make them identifiable as my DOG concept, or as the instantiators of my DOG concept in cognitive processes.[3] I want to stress that the kind of psychological entity I am trying to describe here is compatible with basically any mainstream theory of concepts of any level. For example, if you think that concepts are neutrally instantiated as stores of information that are widely distributed throughout the brain, that is perfectly compatible with concepts being unitary cognitive resources as I conceive of them. The word 'unitary' is not supposed to imply 'necessarily atomic' or 'localised to specific parts of brain tissue', or 'completely static'. It is just meant to imply that the brain

[3] Note that Machery (2009) and Barsalou (2017) use the word 'core' to refer to a putative kind of content that is central and potentially automatically activated when a cognitive process features a given concept. I do not have this notion of 'core' in mind here.

structures that underwrite possession of, say, a DOG concept, are in principle identifiable *as* those brain structures at time *t* and *t+1.*

With this notion in place, I shall outline the received view of concepts in psychological and philosophical research, and show how the received view treats concepts as unitary cognitive resources. I will then argue that this received view fails in principle to provide theories of some 'abstract' concepts, because its fundamental assumptions are incorrect. The received view assumes that for more or less every word of English, we possess a unitary cognitive resource; a concept that corresponds to that word's meaning. I do not argue that there are no such things as concepts or unitary cognitive resources. I simply argue that postulating abstract concepts such as JUSTICE has not helped us to explain phenomena or develop our theories of cognition, and so we do not have any justification for postulating them.

Although the philosophical literature has produced many different views about what concepts are, what they are like, and what they do, an important trend has emerged across theories and proposals. This trend is to talk of concepts as if they were unitary cognitive resources in the sense just outlined. Dove (2014, p. 373) explicitly rejects Fodor's language of thought hypothesis, but he still talks of concepts in a fundamentally Fodorian way, referring to concepts such as DEMOCRACY, ENTROPY, and JUSTICE (using capital letters notation). Carston (2012) uses more explicitly Fodorian language: 'concepts… are constituents of thoughts'. Likewise, Burge (1993) describes concepts as 'the sub-components of thought contents'. Barsalou (2008) proposes that linguistic representations are labels or pointers that trigger the activation of a relevant simulator in cognitive tasks. The 'meaning' of a word is generated by the simulator(s) that is triggered upon encountering that word. Barsalou (2008) states that 'simulators are roughly equivalent to concepts'.

At first glance, Barsalou's view of concepts might seem very different to the others, but for my purposes I don't think this is really the case. Barsalou's view still suggests that there is a correspondence between (English) words and simulators, such that each word is connected to a unitary cognitive resource (or set of such resources). So I think Barsalou's concepts are still unitary cognitive resources, despite the fact that, on his view, concepts *generate* the constituents of thoughts, rather than feature in thoughts themselves. But that isn't really incompatible with the fundamental idea that thoughts can be analytically decomposed into constituent parts, or schematised as a collection of cohesive 'chunks'. Furthermore, these parts or chunks can be traced back to a posited unitary cognitive resource; a concept, and

this concept is in a rough correspondence with an English word. Indeed, Barsalou (1999) himself uses language such as: '[simulators] combine to produce complex perceptual symbols combinatorially'. In the context of that discussion, Barsalou schematises a thought with the content 'the balloon is above the cloud' as involving simulators corresponding to *balloon, above,* and *cloud.* In any case, if Barsalou (2017) would reject the idea that he views concepts as unitary cognitive resources, then that is difficult to square with his talk of such 'concepts' as TRUTH being flexibly 'conceptualised' in specific situations. It seems like TRUTH must be a unitary cognitive resource, or it wouldn't make sense to talk of the *same thing* being conceptualised differently in different situations.

A slightly different picture can be found in Camp (2015). She argues that there is a useful distinction to be made between concepts and what she calls 'characterisations'. On Camp's view, concepts 'function as arbitrary, recombinable, representational bits'. From what I can tell, Camp's view of concepts is roughly in accord with the positions I have just discussed and the notion of unitary cognitive resources. The difference is that Camp identifies a new kind of mental representation, the 'characterisation', that she believes has been overlooked in studies of human cognition. Characterisations are associative, emotionally- and imagistically-charged collections of mental representations that are 'contextually malleable' in such a way as to support and augment the more minimal, language-like (Fodorian), core processing subserved by concepts. For the time being I simply want to point out that despite her distinction, she still describes a theoretical construct that seems like it must be a unitary cognitive resource.

Machery (2009) outlines another anti-orthodox position that is important with regards to the issues I am considering here. Machery suggests that concepts should be thought of as bodies of knowledge that are used 'by default' in higher order cognitive processes. So the DOG concept is the collection of mental representations that features as a matter of course whenever a cognitive process is about dogs. If I understand him correctly, then I think this definition of 'concept' is compatible with my notion of a unitary cognitive resource. Machery (2009, p. 15) also alludes to a connection between words and concepts: 'words seem to be associated with default bodies of knowledge', and he uses capital letters notation to refer to concepts frequently. But Machery spends an entire book (titled *Doing Without Concepts)* arguing that the 'concept' construct is not theoretically useful because the phenomena that it has been used to explain are so diverse that there cannot *be* a category of concepts that has the properties that philosophers and psychologists seem to ascribe

to it. Instead, we should talk in terms of more fine-grained distinctions between empirically-supported constructs such as 'prototype', 'exemplar', and 'theory' (Murphy, 2004). It is important to note here that, in some ways, Machery's overall position regarding concepts is similar to the point I hope to have made by the end of this chapter. However, there are fundamental differences between my view and Machery's. Contra Machery, I believe that concepts are useful constructs in many cases, and that they do good psychological and philosophical work. I am also highly sceptical that the finer-grained notions of mental representations that Machery prefers will deliver us a better account of 'abstract' concepts than the ones we have now. I simply argue against the view that we possess a concept for every word of English (or whatever language), where concept should be construed as a unitary cognitive resource.

Aside from 'what they are', another important aspect of a theory of concepts is 'what they do'. Again, here there is broad agreement across various literatures. Barsalou (1999, p. 383) suggests that concepts 'represent types and tokens … produce categorical inferences [and] combine symbols productively', among other things. Fodor (1998, p. 23) lays out a list of properties that he thinks concepts must have. One especially crucial property is that 'concepts are categories and are routinely employed as such'. So, the DOG concept picks out the category of dogs and not the category of tigers. And in his classic overview of experimental psychology research on concepts, Murphy (2004) spends the entire introduction emphasising the role that concepts play in kind individuation and categorisation. One potential objection to the picture I am painting here is that Fodor means something different when he talks of concepts being 'categories' to what Murphy and others mean when they talk of the role that concepts play in categorisation as a cognitive process. It's possible to read Fodor as making a metaphysical claim about what concepts are, as opposed to a psychological claim about processes that concepts "in the head" feature in, and what behaviour they mediate. However, given that Fodor (1998, 1975) consistently makes a virtue of referring to the psycholinguistic and language acquisition literatures when making his proposals, I don't think this other reading is the correct one. The confusion might rest on Fodor's view about what makes it the case that someone *has* a concept. Fodor argues that, contra to what he calls 'pragmatism', having a concept of X isn't 'having the ability to categorise X'. Instead, having a concept of X is having a symbol in the language of thought that is nomologically locked to X. As far as I can tell though, Murphy et al. could concede this first point and still hold that concepts can (indeed, generally do) feature in the

mental processes that instantiate our categorisations. Furthermore, it also seems to me that Fodor would agree with Murphy et al. on the latter point. So if there is a debate to be had here, I don't think it's relevant for our purposes.

This issue aside, note how the approaches to these two aspects of a theory of concepts – what concepts are and what they do – are pleasingly complementary in some respects. Concepts are the building blocks of thoughts, and they serve to pick out categories in the world. Many of our cognitive processes turn on the ability to make appropriate inferences based on category membership (if that's a tiger, I should run). So what better candidate for enabling these inferences than a unitary cognitive resource that, among other things, represents (or stands for) categories? You and I could almost certainly have a mutually intelligible conversation about dogs. What better way of explaining this ability than the suggestion that we both possess a DOG concept that picks out the same category of things? Note also how, despite the fact that there are substantive differences between theorists concerning the details[4], there is an extraordinary amount of agreement across the philosophical and psychological literatures when it comes down to the basics: It's useful to think of thought as having constituent parts, these constituent parts are (or generate) cognitive stand-ins for categories, the words of natural language are in a rough correspondence with these constituent parts, and it's these constituent parts that a theory of concepts has to provide an account of.

To summarise, I think psychologists and philosophers have in mind the following picture when they investigate concepts. Concepts are (generators of) constituents of thought, and they come with lots of benefits. Perhaps the most important of these benefits is that they enable us to categorise things. Note that this is not to claim that this is all that concepts do, or that concepts 'just are' categories. The way to model thought is to think of it as being made up of, or constructed from, these constituents. A cognitive process that results in my realising or entertaining the proposition that 'I love dogs', therefore, contains the concepts DOG and LOVE, among others. There is a pretty reliable relationship between concepts and words of natural language such that there is roughly a concept in our cognitive repertoire for every word we know (putting synonymy and polysemy aside for the time being). Concepts are unitary cognitive resources in that they are distinctively cohesive and stable. Note that by stable, I do not mean to imply that a psychologist has to believe

[4] This might be putting it somewhat lightly. Fodor's review of Murphy's book ends: 'Gregory Murphy's book tells you most of what there is to the psychology of concepts. Read it, therefore, by all means; but don't even consider believing it.'

that the structure or representational makeup of a concept is immutable. Barsalou makes it very explicit that he believes the opposite. I simply mean that, in principle, it would be possible to 'track' brain structures such that at many different time points and situations, it makes sense to refer to those brain structures as instantiating the same concept. In this picture, it is relatively easy to explain how we acquire our DOG concept and what kind of mental representations underwrite our DOG concept. Although so far it has proven extremely difficult to account for, it is taken as *just obvious* that we also have a unitary cognitive resource corresponding to, say, JUSTICE, and that the structure, acquisition, and format of JUSTICE are among the explanatory targets of a theory of concepts (i.e., unitary cognitive resources). In what follows, I will argue that this is a fundamental mistake.

3.4 Concepts as theoretical posits

In this section, I shall suggest that in at least some cases, the lexemes of English (or any other language) are unreliable indicators of what concepts there are, if we take concepts to be anything like the things that many philosophers and psychologists seem to be talking about. Rendering a word in capital letters in order to distinguish that word from a concept may only give the illusion that there are such concepts in some cases. We should always bear in mind that concepts are posits. They are parts of theories that are supposed to explain something about the mind-brain. The main job of a concept is to explain how it is that humans can cognise about and categorise things. So, the concept DOG – a posit – is supposed to be able to explain how we cognise about and categorise dogs. In the philosophy of science, different attitudes are taken towards the *existence* of theoretical posits. Broadly, two camps emerge: realism and antirealism. The realist doctrine holds that the theoretical posits of our best scientific theories really do exist. The electron is a theoretical posit in physics that figures usefully in the explanations of many physical phenomena. The predictive and explanatory power of the electron is so great that the realist declares that *electrons exist*, even though we do not and could not actually have any direct experience of electrons themselves. Putnam's (1979) 'miracles' argument is a standard realist move: it seems incredible to suppose that it's just a *coincidence* that all of our available evidence is consistent with the existence of the electron. The simplest and most plausible explanation is simply that electrons really do exist.

The opposing view, antirealism, denies this. A popular antirealist argument is that the history of any science is littered with explanatorily and predictively successful theories that then later turned out to be false. For example, phlogiston was posited in

1667 by Johann Becher. Combustible materials contained phlogiston, so the story went, which was released during combustion and absorbed by air. This might sound quaint, but the phlogiston theory actually did quite a good job of explaining certain phenomena observed in combustion. But no one today believes in phlogiston. The antirealist then points out that we have no reason to expect that we have miraculously discovered *the* correct theoretical posits at any given point in time: who is to say that in the year 2200, the theory of the schmelectron, a posit that provides an explanatory and predictive advantage over the old electron theory, will not be taught in secondary schools around the world? And so the antirealist concludes that it isn't safe to assume that any theoretical posit 'really exists'. But importantly, the antirealist can still accept the explanatory and predictive power of successful theories. So the antirealist might say that the universe behaves *as if* the electron existed. And for most purposes, that's good enough.

I am bringing all of this up because, although they seem diametrically opposed, the realist and the antirealist actually agree about one crucial issue, and I think a lot turns on this issue when it comes to theorising about concepts. The issue that the realist and the antirealist agree on is that *we should only entertain posits that have explanatory value.* By this I just mean that theoretical posits need to help us predict or explain the phenomena that we are interested in explaining. The contemporary realist and the antirealist *agree* that phlogiston (probably) does not exist, and that there are now more successful theories of combustion that do not include it. The point at which the realist and the antirealist start debating is the point at which a theory is explanatorily successful, or a theoretical posit has explanatory value. My aim over the next few sections is to argue that although *some* proposed concepts (theoretical posits) have explanatory value, and potentially belong in theories of concepts, some do not. By this I mean that some concepts (e.g. DOG) play a clear role in explanations of human behaviour and cognition, whereas other concepts (E.G JUSTICE) do not have a clear role at all; in fact, our theories of concepts work just the same if we remove JUSTICE from them. I hope to demonstrate a methodology for identifying good candidates for concept-hood, and bad candidates. I will suggest that there are proposed concepts that are extremely difficult for any theory to account for, and the reason for this is these proposed concepts might not be concepts at all. I will try to show that, although it *is* useful to posit concepts such as DOG, it might not be useful to posit a broad range of what have been traditionally referred to as 'abstract concepts'. Experimental psychologists, and at least some philosophers, believe that cognitive science really is a science. If that's true, and I

agree that it is, then we should treat cognitive science as such, and only posit constructs that are useful. I focus on the alleged concept JUSTICE in order to make my points, although I suspect the same arguments could be made generally against postulating a wide range of 'abstract' concepts. I think that it would be much more useful to examine certain human cognitive capacities without restricting ourselves with the impossible requirement that the explanation of these cognitive capacities must be subsumed under a single label in capital letters.

I am not the first person to note that there are a range of 'concepts' that have proven particularly difficult to analyse on any theory of conceptual content (Barsalou et al., 2008; Crutch and Ridgway, 2012; Dove, 2016; Kousta et al., 2011). These are the 'abstract' concepts such as JUSTICE, ENTROPY, and DEMOCRACY. I argue that the reason these concepts are so difficult to analyse is because they are not plausible candidates for concept-hood in the first place. I claim that there is no such thing as a building block of thought, JUSTICE, that features in all or even most thoughts 'about justice'. Nor is there a neural simulator that generates all or only justice simulations. There is no such thing as a unitary cognitive resource that plays a role in justice categorisations, or events that involve justice-or-an-aspect-thereof. 'Justice' is a word that can be used to mean any number of things. But we should not let that confuse us into thinking that there is such a thing as JUSTICE. To make this argument, I am first going to consider the case of DOG, and show how it is extremely useful to posit DOG as part of an explanation of how we categorise and cognise about dogs. Then I am going to argue that it *isn't* useful to posit JUSTICE as part of an explanation of how we categorise and cognise about anything. So *even a hard-line antirealist* shouldn't think of JUSTICE as being a concept in the first place, because JUSTICE is not a useful posit. And concepts, after all, are 'just' posits.

Let us consider the reasons we might want to posit a DOG concept. It is not controversial that, having acquired the appropriate sensorimotor experience, we would expect most humans to be quite proficient at recognising certain medium-sized objects as being members of the set of dogs. With the appropriate training in a given language, most humans would also not have much trouble in using the appropriate word to pick out this set. It's extremely likely that any given human would be able to report having 'thoughts about dogs', and we would understand them to be communicating the idea that the object of these thoughts was (members of) this set. These facts are obviously in need of explanation. The almost universally-accepted explanation goes like this: we posit that human mind-brains have these things called concepts. Concepts are the fundamental units of cognition. Concepts function to pick

out categories, and they feature in neurocognitive processes in such a way as to explain *why* neurocognitive processes produce the results that they do. Ultimately, concepts will be instantiated somehow by our neural circuitry, although exactly how this happens is at present something of a mystery. DOG is one such concept. Possessing a DOG concept allows someone to categorise and cognise about dogs as such. In most cases, acquiring a DOG concept is largely a matter of gaining the appropriate sensorimotor experience of dogs and somehow storing this experience in the appropriate way (or, if you are a Fodorian, it is entirely a matter of your mind-brain having a symbol that stands in the appropriate causal relations to dogs).

This is an extremely compelling explanation, and I think something along these lines will produce a powerful psychological theory of concepts (and hence, cognition). However, and I stress this again, it is crucial to keep in mind that despite how compelling this explanation is, DOG is *just a posit*. We have observed that human beings respond in reliable ways to the presence of dogs, and in order to explain this (mental) behaviour (on a loose understanding of what constitutes behaviour), we posit the DOG concept. One of these (mental) behaviours might be to realise that there is a dog in the room. It is useful to use the posit, DOG, to talk about the neurophysiological responses, patterns of brain activation, cognitive processes, and behaviour that underwrite this realisation. DOG is just a way of thinking about thought in such a way as to aid our explanations and allow us to make (hopefully accurate) predictions about human behaviour and descriptions of cognitive processes. The electron has turned out to be an extremely useful posit for explaining many physical phenomena. It is so useful that we might even be tempted to say that electrons exist. The ultimate aim of psychology and neuroscience is to provide an explanation of human behaviour and cognitive processes. With regards to these aims, positing a unitary cognitive resource that serves to pick out the category of dogs has explanatory value, and it could be that you believe that DOG is so crucial to psychological and behavioural explanations that we should believe that it exists. But that is all DOG is: it is a posit with explanatory value. The question I will now consider is whether JUSTICE is a posit with explanatory value. I will argue that the answer to this question is ‘no’.

3.5 JUSTICE in a language of thought

Let’s consider some popular answers to the question of what properties DOG and JUSTICE are supposed to have, such that they can play an explanatory role in behaviour and cognitive processes. I want to show that all of these answers and

explanations make sense in the case of DOG, but make no sense in the case of JUSTICE. As we have seen, one traditionally popular answer is that DOG and JUSTICE are symbols in the language of thought (Fodor, 1998, 1975). On Fodor's view, concepts are individuated according to the causal relations that hold between them and the mind-external world. So DOG is a symbol that stands for, features in inferences about, and categorises dogs, because it stands in certain causal relations to dogs 'out there' in the world. A tokening of the DOG symbol is caused by dogs, or is caused because it stands in the appropriate relations to other symbols that have themselves been tokened. The DOG symbol itself is atomic: it cannot be decomposed or analysed into constituent parts. It has no structure. Now, consider JUSTICE. On Fodor's view, if JUSTICE is a concept, then JUSTICE must also be individuated by its causal relations with the outside world. But here we come up against an immediate challenge. It is extremely difficult to give a satisfying account of what it *is* that JUSTICE stands in causal relations to. What is the category of justices? What *are* the things 'out there' that cause my JUSTICE concept to be tokened? One response here could be to describe some situations that seem to involve some element of justice, and look for some commonalities between them. Perhaps those commonalities are the things that stand in the appropriate causal relations to JUSTICE. Here are two such situations. Given the publicity constraint, we shouldn't have to worry about whether the same concept could be tokened in each case (or, at least, a highly similar concept if you think publicity is too strong a constraint).

A. Suppose a man who believes his chickens to have been waylaid successfully accuses someone of the crime, and this unfortunate ends up in the pillory. Up until 1837, being locked in a pillory was a common punishment in England for various criminal offences. Upon witnessing the public humiliation and physical violence that invariably followed, and perhaps even joining in himself, the accuser is satisfied that justice has been done.

B. Suppose that a world-class athlete stands accused of taking prohibited substances that regulatory bodies believe to be both dangerous and likely to convey an unfair competitive advantage over those who have not taken them. But, after much diligent work from the athlete's lawyer, it transpires that there was no possible way that the athlete *could* have taken the substances, and that their test results had been confused with someone else's. After the jubilant press conference that follows this announcement, the athlete's lawyer is satisfied that justice has been done.

One obvious commonality here is that the rule of law has been carried out in both situations. But it is quite clear that JUSTICE cannot just be 'that mental representation which is tokened when an agent understands the rule of law to have been carried out'. (Let me stress emphatically that I am talking about JUSTICE here, and not justice. I do not have anything definitive to say about what justice is: I take it as a virtue of my position that I can avoid having to do so). For one thing, I find it quite likely that any number of people reading this would be reluctant to allow that the alleged chicken thief's treatment was just, but they also would not deny that the rule of law had been carried out (or what passed for the rule of law in medieval England). And so their JUSTICE concept must stand in causal relations to mind-external entities over and above carryings-out of the rule of law. It is also not hard to come up with situations in which the rule of law has been carried out and yet most people would take them to be *manifestly* unjust. Take, for example, any of the cases in which someone has been punished by a judicial system who was in fact innocent. Cases where a judicial system convicts an innocent party are not cases in which the rule of law has not been carried out: it was the carrying out of the rule of law that secured the guilty verdict in the first place (or at least, that is the assumption that most judicial systems are predicated on). Perhaps we should say instead that JUSTICE is 'that mental representation which is tokened when an agent understands the rule of law to have been carried out, believes that an accused is in fact guilty, and also believes that the accused deserved whatever it is that ultimately happened to them'. Unfortunately, it isn't difficult to come up with scenarios in which JUSTICE is apparently not *only* tokened by the punishment or conviction of a guilty party. Witness the athlete's exoneration. Perhaps we could modify our account of what JUSTICE stands in causal relations to with the addition of a clause to the effect that JUSTICE can also be tokened by situations in which it is discovered that an accused is innocent and subsequently exonerated.

The problem here is that, although we have gone some way towards a sketch of when it is that our JUSTICE concept is tokened, we still have not really explained anything about human behaviour or mental processes. We are no closer to an explanation of how JUSTICE is acquired, or what it's made of, or *how* it has the causal powers that it supposedly has. All we have done is provide a non-exhaustive list of situations, and stipulated that they cause a tokening of JUSTICE (note the capitals). I think it's clear that this does not count as an explanation of cognition or behaviour. Also, if concepts are individuated by their causal-relations to mind-external entities, then it doesn't seem unreasonable to enquire what those mind external entities might

be, or maybe how we could go about finding out. It's important to note here that Fodor (1998) was quite clear that a lot of concepts are innate, and that basic level concepts aren't amenable to structural analysis anyway. So *he* probably isn't bothered too much by failing to meet these explanatory targets. But as we saw in the introductory paragraphs of this chapter, most of recent psychology rejects both of these ideas while still acknowledging some of Fodor's considerations. So a modern psychologist who endorses a Fodor-esque view really *does* have to explain how JUSTICE is acquired, what causal powers it has, and why it has these causal powers. Notice that in the case of DOG, these things more or less come for free: having DOG is just having a symbol in the language of thought that is nomologically locked to dogs. Having DOG explains how you can categorise dogs, because incoming dog-like sensorimotor input triggers DOG (by virtue of DOG's law-like connection with dogs).

And even if this didn't worry us, it is not hard to come up with counterexamples that suggest that we cannot have exhausted the things that JUSTICE supposedly stands in causal relations to. It is also perfectly felicitous to describe as 'just' a situation in which a piece of cake is sliced evenly so that two equally-deserving children each get the same amount. I think the problem here is that any attempt to specify some properties of situations 'out there' that stand in the appropriate causal relations to a mental symbol such that that symbol is JUSTICE will ultimately turn into attempts to define justice. But as Fodor (1975) argued at length, whatever they are, basic/lexical concepts can't be definitions, and it is also extremely difficult to define anything, let alone 'what justice *is*'. My point is ultimately totally banal: the word 'justice' can be used to talk intelligibly about an infinite number and variety of situations. But I think that there is an important moral to drawn. We want to explain the human (mental) behaviour that accompanies situations which we might describe by using the word 'justice'. But we are not going to get very far by positing a unitary cognitive resource that features in all of the cognition and behaviour associated with every conceivable way in which the word 'justice' might be used. I think it is more likely that understanding different uses of the word 'justice' in different situations, or being tempted to communicate in two different situations by using that word, may involve sets of mental representations and cognitive capacities that are more or less distinct from each other, and that none of these sets plausibly corresponds exactly and only to the posit, JUSTICE. And, as we shall see later, these potentially distinct mental representations and cognitive capacities explain human (mental) behaviour independently of being subsumed under some JUSTICE concept.

Furthermore, it could be that, far from furnishing us with explanations and predictions, positing JUSTICE might actually make those goals harder to achieve. The aim of psychology is to explain and understand human behaviour and cognitive processes. Imagine that you perceive a fitting reward to have been finally bestowed upon a long-deserving and hitherto unrecognised friend. I suppose we might allow that justice had been done in some sense: your friend has finally been acknowledged as the hero she is. This shouldn't be controversial, because if you find the previous sentence intelligible at all, then you must have understood my use of the word 'justice' at least partially. You might wonder and desire to explain:

1. What mental representations, brain structures, and processes underwrite my ability to perceive my friend's reward, my disposition to want her rewarded, and my taking satisfaction in it?
2. What mental representations, brain structures, and processes underwrite my feeling that there was something not quite right about my friend's efforts being unrecognised?
3. How did these mental representations, brain structures, and processes come to be – how did I acquire them?

If you assume that a JUSTICE concept is going to figure usefully in responses to some of these questions, then a Fodor-style answer is that an unanalysable, atomic unit of the language of thought stands in the appropriate causal relations to entities in the outside world such that it is tokened by the act of your friend receiving her reward. *And* that the same unanalysable, atomic unit of the language of thought is also tokened whenever you understand a guilty party to have been convicted of a crime. As to *how* this symbol underwrites these cognitive processes, and what properties it has that explains how this symbol underwrites cognitive processes occurring in seemingly disparate situations, and producing seemingly disparate effects, a Fodor-style theory is silent. Basically, in order to explain how human beings behave and cognise in all of the different situations that we might describe using the word 'justice', a Fodor-style theory can only stipulate that the JUSTICE symbol stands in the appropriate causal relations to mind-external entities, and enters into the right syntactic relations with other concepts. However, it doesn't seem like we're going to be able to say anything about what these causal or syntactic relations are, and so that's the end of the story.

But this is contrary to intuition and evidence. It does seem to be possible to investigate the questions posed in the previous paragraph, although for the time being these investigations are surely preliminary. For example, there exist empirical investigations of the neural structures implicated in perceiving the rewards of others (Lockwood et al., 2015). There exist empirical investigations of children's understanding of what constitutes fair behaviour (Wittig et al., 2013). And there is at least one large and influential body of theoretical work that attempts to explain how humans mentally represent the cognitive states of other human beings, and how they act on the basis of those representations (Sperber and Wilson, 1995). This work, and other work like it, seems like a plausible basis on which to start formulating explanations of human behaviour and cognition in *some* of the situations that we describe by using the word 'justice' (and a great many other words). But these explanations do not require us to posit JUSTICE. One response here is to argue that other concepts in a language of thought are prerequisites for having JUSTICE, and that with these prerequisites in place, JUSTICE does more explanatory work. For example, you might suppose that under the right conditions, someone who has the concepts, REWARD, PUNISHMENT, and FAIRNESS can acquire JUSTICE. And *then,* JUSTICE can have causal powers partly in virtue of these prerequisite concepts. I have two counterarguments to this kind of move.

First, it's arguably un-Fodorian: Fodor held that concepts are not individuated by their relationships with any other concepts, but solely by their nomological locking with something external to the mind. The reason that Fodor rejected the idea that concepts can be individuated by their relationships with other concepts was because he thought that if that were true, then publicity would be violated. The reason that publicity would be violated is that, if some concepts are individuated by things "in the head", then there is scope for concepts being individuated differently for different people (because what is "in the head" may vary interpersonally). Now, you might think that Fodor was wrong about that argument, and/or you might not think publicity is important. In that case, my second counterargument is that the prerequisite concepts REWARD, PUNISHMENT, and FAIRNESS do all of their explanatory work independently of whether you stipulate that they have something to do with JUSTICE. Suppose for a second that there are such things as mind-external properties that the mind-brain becomes nomologically locked to, such that we would want to say that these mind-external properties individuate the concept REWARD. (And, suppose we could say what those properties are). Any agent who possess REWARD has all the resources they need to distinguish this category, and have thoughts that feature the

concept REWARD. But what we want to know is what the mind-external properties that individuate *JUSTICE* are, and where *JUSTICE's* causal powers come from. Speaking very crudely, suggesting that JUSTICE is the concept that equals REWARD, plus PUNISHMENT, plus FAIRNESS, is not going to give us any explanatory power that we didn't already have. Our explanatory power is going to just be whatever explanatory powers that REWARD, PUNISHMENT, and FAIRNESS provided us. There is nothing to be gained from stipulating that JUSTICE will somehow fall out of these prerequisite concepts, unless we can come up with some properties of JUSTICE itself.

Instead, I think that it should be possible to formulate these descriptions and explanations without positing a unitary cognitive resource that potentially corresponds to every way in which the word 'justice' was used in all of the examples given above, and we should be free from the restriction that the same unitary cognitive resource has to play a role in all or even most situations we describe by using the word. I claim that there is no such unitary cognitive resource: attempts to spell out how JUSTICE is acquired or represented, or what kinds of representations it is made of, are futile in principle because it is not useful to posit such an entity. Psychologists do not need it in order to investigate any of the phenomena they might be interested in. Instead, psychologists could (and do) focus on more tractable questions, such as, 'at what age do children move beyond the understanding that 'fair' always means 'equal' no matter the circumstances?', or 'is there a neural correlate of the extent to which people take pleasure in the success/failure of others'? *Depending on what you mean by the word 'justice' in a particular case*, answers to these questions might tell us something about human cognition and behaviour in that case.

The word 'justice' can be used to communicate about many things, but there is no unitary cognitive resource that can usefully figure in explanations of what mental representations underwrite our understanding of all these things. The fact that we can use the word 'justice' to communicate about both an athlete's legal exoneration and an egalitarian distribution of cake should not suggest that we deploy the *same* unitary cognitive resource(s) should we find ourselves in both situations. It should simply suggest that we can use the English word 'justice' to communicate about a variety of things (I examine this issue in detail in Chapter 8). The notion that both sharing cake, and acquiring the understanding that an athlete has been cleared of a crime, depend on the *same* unitary cognitive resource should require a lot of evidence, but in the psychological and philosophical literatures it seems to have been assumed.

Furthermore, no evidence of this sort has been provided by psycholinguistic experiments that feature the word ‘justice’.

3.6 JUSTICE in a network of simulators

I have just argued that a classic Fodor-style approach to concepts will not work when it comes to explanations of certain human behaviour and mental processes that we *assumed* must involve the concept JUSTICE. But what about a Barsalou-ian simulator theory? As we have seen, it is relatively straightforward to see how DOG might be some structured neuropsychologically-instantiated entity composed (mainly) of sensorimotor representations, and how this structure and makeup might explain our ability to categorise and cognise about dogs. Now, what are the mental representations that JUSTICE is composed of such that it allows us to cognise about justice (or situations involving an element thereof)? It’s clear that, narrowly construed, sensorimotor representations on their own won’t be able to do the job of constituting JUSTICE. We might associate a certain kind of black robe and a white wig with a situation that could be felicitously described using the word ‘justice’. But I don’t think we would want to say that justice (without capitals) has anything much to do with black robes and white wigs in and of itself, and sensorimotor representations derived from perceptual experience of black robes and white wigs do not seem to be useful in explaining why it is that I might believe sharing cake to be just, or how I acquired the mental representations that underwrite that belief. However, as Barsalou (2016) has recently pointed out, typically, embodied cognition theories aren’t actually committed to the claim that cognition and categorisation has to be explained purely in terms of sensorimotor representations even if we take care to construe the term widely. Barsalou also proposes that along with simulators, we possess another kind of cognitive structure: skeleton blueprints of objects and events called *frames*.

Barsalou’s frames are an often-overlooked aspect of his theory, and they seem like useful constructs for meeting our explanatory aims regarding some situations we might describe by using the word ‘justice’. However, I think that these frames do their explanatory work regardless of whether we stipulate that they are part of an overall concept labelled in capital letters (JUSTICE), and so there is no need to make this stipulation. Barsalou (1999, p. 590) describes frames as blueprints for how situations and events are constructed, and how parts of an object come together to form a whole: ‘a frame is an integrated system of perceptual symbols that is used to construct specific simulations of a category’. We might have a frame for any object or

event we experience. A going-to-the-shop frame might have slots for a clerk, a customer, and a product. Simulators could generate largely sensorimotoric/affective/proprioceptive simulations that fit neatly into these slots, and the frame provides a way of organising and structuring these simulations in such a way as to facilitate task-appropriate behaviour and mental processes. Let's apply this idea to JUSTICE. Perhaps we have a frame that corresponds to 'courtroom-setting-in-the-United-Kingdom'. This frame might have all sorts of slots in all sorts of relations to each other. One of these slots might be for a person wearing a white wig and black robe who stands at the top of the social hierarchy in the context of the frame. Another slot might be for a largely stationary official-looking person who occasionally responds to instructions from the person in the wig, and so on. Suppose that an array of simulators could produce simulations to fit in these slots in such a way as to organise a thought expressible by an utterance of "justice has been done as a result of a guilty verdict". I think that positing such a set of mental representations is just as reasonable as positing DOG, especially for someone who for whatever reason finds themselves in U.K. courtrooms frequently.

The problem for Barsalou's account arises when it comes to specifying what the relationship is between this courtroom frame and JUSTICE. One option is to say that this courtroom frame is part of the makeup of an overall JUSTICE concept (or simulator, as he would call it). I think this is a mistake, because this strategy forces us into the position that there must be something in virtue of which a UK-courtroom-setting frame belongs to the same unitary cognitive resource as a simulation of an experience of sharing cake. Otherwise, we would have no theoretically motivated reason to want to subsume them under the same JUSTICE concept. Why shouldn't we say instead that I have a SHARING CAKE simulator/frame, and a UK-COURTROOM simulator/frame? Positing these mental representations can explain how I can think about UK courtrooms and recognise being in a situation in which cake is shared, but there is no need to posit JUSTICE to explain these capabilities. And in any case, the problem that faced the classic Fodor-style account comes back in full force. The answer to the question of what mental processes and behaviours JUSTICE plays a role in becomes, "any and all of the processes and behaviours associated with any and all of the situations that we can communicate about by using the word 'justice'". The task of specifying what category JUSTICE corresponds to then becomes very difficult indeed, because there doesn't seem to be a principled way of defining what that category is.

The second option is to accept that if this courtroom frame (or cake-sharing frame, for that matter) has explanatory value, then it has this explanatory value quite independently of a connection to a JUSTICE concept. But if you believe that there *is* such a thing as a JUSTICE concept, and it seems to me that Barsalou does (see Barsalou and Weimer-Hastings (2005) especially) then we are again no closer to specifying how that concept was acquired, what it's made of, what behaviour and mental processes it plays a role in, or how it plays this role. Indeed, there does not seem to be a role for JUSTICE to play. So there is nothing to be gained from maintaining that the courtroom frame is part of a JUSTICE concept, because the courtroom frame itself can explain how people who possess it can recognise courtrooms in the United Kingdom, and behave appropriately when they find themselves in one, and form the appropriate beliefs and come to the appropriate realisations when things happen in one. We don't need to embed this courtroom frame in anything else in order to explain these behaviours and cognitive processes, and there just doesn't seem to be anything to be gained from stipulating that this courtroom frame is part of an overall JUSTICE concept. We haven't explained anything about the acquisition of JUSTICE or what mental representations JUSTICE consists of, or how JUSTICE mediates cognitive processes and behaviour. But we *might* have come up with a hypothesis about how human beings acquire concepts that mediate being in a U.K courtroom setting, what mental representations those concepts consists of, and how *those* concepts mediate cognitive processes and behaviour. And to me that seems like a promising achievement rather than defeatism. Note here that I am not committing to anything about the theoretical status of these particular courtroom posits: my point is that to the extent that they do turn out to have explanatory value, we should be happy to include them in our psychological theories.

I want to end this section with a thought experiment. It could be that you are still sceptical that we could get by without positing JUSTICE despite the problems discussed in the preceding paragraphs. Perhaps it still seems like there really is some unitary cognitive resource that the word 'justice' picks out, and even though it is very difficult to couch our explanations in terms of it, this cognitive resource might figure in behaviour and mental processes as a matter of fact. This thought experiment does not constitute a knockdown argument, but I hope it does go some way to shoring up my position against this kind of intuition:

Imagine that at some point in the future we invent a ray gun that is capable of 'knocking out' specific concepts. The rest of the cognitive system is left totally undisturbed, and *only* that circumscribed set of mental representations hypothesised

to instantiate a particular concept are 'deleted', as it were, from someone's conceptual repertoire.[5] Suppose we set the dial to the notch labelled 'dog', and pull the trigger at our test subject. What specific capacities and cognitive processes would we interrupt? Granting that the ray gun's powers are genuine, relatively straightforward answers suggest themselves.[6] If we placed a dog in front of our test subject, they would not be able to recognise it; perhaps they would ask us what this strange new animal was. If we ray-gunned many subjects and asked them: "There's a rare animal called a 'dog' - do you think it is a four-legged mammal, or a small biting insect?", perhaps they would answer at chance levels (although would they recognise that they had heard the sound [dɒg] before?). But they would still be able to recognise a cat, if we placed one of those in front of them instead. They would correctly answer questions like, "is cow's milk a staple food in some human cultures?", and they would probably run from a rampaging rhinoceros instead of offering to play chess with it. But what would happen if we turned the notch to the dial labelled 'justice'? Would the subject lose the inclination to share anything ever again? Would they be unable to understand what the word 'punishment' is supposed to refer to? Would they lose all sense of law and order? Would it become apparent that they lacked a sense of empathy? Or all of these things? Or is there a specific subset that would be affected, and could we predict with any confidence which subset it would be? It could be that intuitions about this thought experiment will vary, but my own response is, unsurprisingly, that it is entirely unclear what effects the ray gun would have on the subject if we tried to knock their JUSTICE concept out. And I think this is again indicative of the vacuity of JUSTICE as a theoretical posit.

[5] If you are a holist who believes that concepts are individuated by their relations to all other concepts in the cognitive system, then I think you would probably object that this is impossible in principle. In that case, imagine that the ray gun maximally destroys a target concept while minimally altering all other concepts.

[6] The reader might wonder if studies of Semantic Dementia (SD) bear on this thought experiment (Bonner et al., 2009). SD patients exhibit a specific impairment of 'semantic' memory: they lose their memories of what objects *are*, as well as what the referents of words are. Typically, this pattern of deficits is presented as a loss of 'conceptual' knowledge. However, SD is a rare, hard-to-investigate disease; stimulus sets in experiments are small; the number of participants in experiments are also small; and conflicting findings are common. Moreover, studies of SD patients that reference the concrete-abstract distinction use linguistic stimuli and responses to draw conclusions about the conceptual system as a whole. It isn't clear that losing the ability to recognize the word 'justice' entails the loss of all of the behaviors and cognitive processes that JUSTICE is supposed to feature in. For these reasons, it is probably premature to draw any firm conclusions about sufferers of SD. In any case, the SD patient is unlike our ray-gun victim in an important respect. SD patients gradually lose whole swathes of their conceptual repertoires, and are not impaired on 'just' one item, so we could not use them to investigate what would happen if a single concept is lost from someone's conceptual repertoire.

To summarise: 'justice' is a word that we can use to try and communicate various things. If we want to learn about how human beings respond to moral dilemmas, societal rules, perceived crimes and punishments; how children come to form beliefs about group behaviour; what mental representations cause or constitute our desire that a certain outcome obtain; and so on, then I think doing the relevant experiments and sociological investigations will tell us a lot. They will tell us about some cognitive processes involved in some of the things that we might try to communicate about when we use the word 'justice'. But they won't tell us about JUSTICE. JUSTICE is just a posit. It is part of a model that we use to understand people. But we should only include it in our model if it is actually helpful to do so. And the upshot of the preceding discussion is that it is not helpful to include JUSTICE in our model of how the human mind-brain works. So JUSTICE is not an abstract concept – it isn't a concept at all, because there is no theoretically coherent mental particular or behaviour that corresponds to it. Instead, it seems more useful to posit other unitary cognitive resources that figure in specific human behaviours and cognitive processes, such as deciding to share a lucrative resource, or taking satisfaction in the announcement of a punishment, and so on. The behaviours and thoughts that occur in situations that we describe by using the word 'justice' aren't mediated by a JUSTICE concept. The mental representations and brain structures that underwrite these behaviours and cognitive processes do not correspond to a single cohesive construct whose organisation and composition will explain such a diverse and complicated array of human thoughts and acts. The words of English are not always good indicators of what unitary cognitive resources we have.

3.7 Consequences of giving up the concept JUSTICE

At first glance, it may seem like the position that I am advancing here is incompatible with almost every theory of thought and/or concepts there is. This would be a consequence so negative that it might be sufficient grounds on its own to reject my position. However, I believe that the most popular theories of thought are actually compatible with my position. The only requirement of these theories that needs to be relaxed is the requirement that words are always useful in identifying what concepts (i.e., unitary cognitive resources) there are. So, it could be that thought is language-like in Fodor's sense, and that the units over which cognition operates are atomic, arbitrary, syntactically re-combinable discrete elements. The only modification that my position requires is accepting that sometimes words might identify a discrete

element of this theory, but sometimes they don't. Likewise, it could be that thought is the product of dynamic interactions between simulators of sensorimotor experience and frames (Barsalou et al., 2008). The only modification that my position requires is that sometimes words might identify a cohesive set of such simulators and frames, but sometimes they don't. So currently-popular theories of how the mind-brain works are actually mostly untroubled by the arguments I have been making.

However, there is one important and very unpleasant consequence of my arguments. If I am at all correct, then I think it is clear that the experimental techniques employed in neuropsychological investigations of 'conceptual processing' will not tell us anything about JUSTICE, or the difference between concrete and abstract concepts generally. Flashing up the word 'justice' on a computer screen and measuring how long it takes a participant to decide that the letter string <justice> corresponds to an English word, as happens very often in psycholinguistic experiments, is unlikely to be informative at all about the kinds of cognitive processes involved when human beings engage in and make decisions about those situations and relations that we might refer to by using the word 'justice'. This is because we can use the word 'justice' to talk about an infinite number of things and, as I have been trying to argue, there doesn't seem to be any one thing that corresponds to a conceptual 'core' that unites all or even most of these things. However, this is precisely the kind of evidence that psychologists and psychologically-minded philosophers have relied upon when constructing their theories about what JUSTICE might be and what properties it has that separate it from DOG. That is to say, this is the kind of evidence that is used to support and investigate the distinction between concrete and abstract concepts. So the extremely unpleasant consequence of the argument I have been making so far is that empirical psycholinguistic work will not help us investigate what *some* 'concepts' are or what properties they have, because in some cases there is nothing to investigate. We do not possess a unitary cognitive resource that corresponds to the word 'justice', and so flashing the word 'justice' up on a computer screen cannot tell us about that resource.

Before acknowledging some obvious objections to the arguments I have just presented, I want to spend some time trying to convince you that instead of being completely destructive, the position I am advocating comes with some substantial benefits. My position might seem destructive because historically, concreteness has been such an important psycholinguistic variable, and there have been many attempts to try and incorporate abstract concepts into our theories (Barsalou and Wiemer-Hastings, 2005; Borghi et al., 2017; Dove, 2016; Kousta et al., 2011; Löhr,

2017; Troche et al., 2014). If I am right, then the utility of a large body of experimental work is in question, and some of these attempts might have been for naught. As I noted above, many others have pointed out that abstract concepts in general pose large problems for our understanding of concepts and cognition (Barsalou and Wiemer-Hastings, 2005; Crutch and Ridgway, 2012; Hamilton and Coslett, 2008; Wiemer-Hastings and Xu, 2005). Traditionally, the assumption has been that if you endorse an embodied cognition account, these problems are especially difficult (Goldinger et al., 2016; Mahon and Caramazza, 2008). The key insight of embodied cognition approaches is that cognition is importantly continuous with perception and action, and that this idea plays a hugely useful role in explanations of cognitive processes and experimental data (de Vega et al., 2008). The problem that abstract concepts pose is that they are by (negative) definition fundamentally non-perceptual, and non-actional. Theories of cognition that appeal to perception and action mechanisms in order to provide accounts of conceptual processing therefore have an especially difficult challenge if they want to account for these concepts. The standard conclusion of the amodalist is that embodied cognition must be false for this reason (Goldinger et al., 2016; Mahon and Caramazza, 2008). However, as Barsalou (2016) and Prinz (2004) have highlighted, the problems caused by abstract concepts are just as deadly for the amodalist. Simply stipulating that "concepts are amodal" does not help us provide a compelling account of abstract concepts: simply using the *label* 'amodal' does not count as an explanation of how we acquire JUSTICE, or other alleged abstract concepts. As I argued above, a classic amodal account like Fodor's (1998, 1975) does more or less as well with DOG as embodied cognition accounts, and just as poorly with JUSTICE. So if I am right, then I think *all* accounts of concepts are in a better shape than they were before.

If we really can't provide satisfying explanations of the acquisition of JUSTICE; or the role it plays in cognitive processes and behaviour; or the mental representations that constitute it, then what I take myself to be providing is a principled way of avoiding these problems while safeguarding the explanatory success we have already achieved. Recall that one of my arguments against the existence of JUSTICE in a Barsalou-ian simulator theory was that there are other posits that could plausibly explain the behaviour and cognition we wanted to explain, such as a cake-sharing frame. It seems to me that the cake-sharing frame is a perfect fit with the rest of embodied cognition's commitments. The cake-sharing frame is plausibly acquired on the basis of sensorimotor, affective, and proprioceptive experience of sharing cakes, and it is in virtue of this that the cake-sharing frame could play a role in mediating

thoughts and behaviours in situations in which I share cake. For this reason it does not pose the same problems that JUSTICE does. Likewise, in Fodor's case, we could not tell a satisfying story about what the individuation conditions for JUSTICE are. However, I think we *could* tell a satisfying story about what the individuation conditions for CAKE-SHARING are, or at the very least this task will be easier than in the case of JUSTICE.

Now that I have laid out my case, I want to head off two potential misunderstandings of the view I am trying to convince you of. I have argued that JUSTICE is not a useful posit, because it doesn't actually help us to explain any behaviour or cognition. Because featuring in explanations of behaviour and/or cognition is the main job of a concept, I conclude that there is no such concept, JUSTICE. The first potential misunderstanding of my view is to confuse it with the claim that there is no such thing as justice (note the lack of capitals); or with the claim that human beings do not have a sense of right or wrong; or that there is no such thing as morality or moral behaviour, and so on. But this is absolutely not what I am arguing. Of course most humans have a sense of right and wrong, and most humans have strong beliefs about what counts as moral behaviour, and most humans have a view about what the right order of things in society is. I am not contesting these facts. Rather, I am arguing that if we want to provide a *neuropsychological explanation* of these facts, and hopefully cognitive science will be able to shed some light on them, then positing JUSTICE does not help. Instead, we should posit, and investigate, other concepts that mediate specific cognitive processes and behaviours (for example, a frame that organises thoughts and behaviour in situations in which we share things). There is a temptation to say that if we sum up all of these more specific concepts, then perhaps a JUSTICE concept will fall out of them somehow. However, I think this is a mistake: our theories do not gain any explanatory power if we stipulate that there is a JUSTICE concept that subsumes more specific concepts. Instead, the more specific concepts do all the explanatory work they are supposed to do independently of any alleged connection with a JUSTICE concept. The second potential misunderstanding is to confuse the technical sense in which I am using the phrase, 'the concept JUSTICE' with other uses of the phrase 'the concept justice'. In the sense in which I am using this phrase, JUSTICE is an alleged unitary cognitive resource and it features *only* in theories and models of the human conceptual system. It is a component of a psychological explanation of the human mind. I am *not* arguing about 'concepts of justice' as they may apply in ethics, law, sociology, theology, or any other discipline. Researchers in these fields may find immense value in couching their

investigations in terms of concepts of justice, and work in these fields *may* inform what we might want to say about some unitary cognitive resources in the domain of psychology. But I am arguing *only* about something "in the mind", or at least in our models of the mind, and it is crucial to recognise this. One way to illustrate what I mean by this is to consider the following two sentences:

(1) Humans do not possess the concept JUSTICE.
(2) Humans have a sense of justice.

If my view is correct, then the statement expressed in (1) is true. But depending on what a speaker means, the statement expressed in (2) may well be true as well. That is, there is no contradiction here.

I have spent a very long time focusing on JUSTICE as an example of an abstract concept par excellence. However, I don't take myself to *only* be arguing that JUSTICE is not a concept. I think we could use the same strategy I have employed here to eliminate other alleged concepts that cause trouble for our theories of cognition. In this chapter, I have applied this methodology to DOG and JUSTICE. DOG, I think, has passed the test, and we should be happy to include DOG as an item in our theories of concepts. JUSTICE has not passed the test; JUSTICE should not be included as an item in our theories of concepts. The test is to assess whether an alleged concept actually helps us meet our explanatory and predictive aims in an account of human behaviour and cognition. There are a host of other capitalised words to be found throughout the psychological and philosophical literatures that are supposed to pick out concepts that are theoretically difficult and 'abstract' (DEMOCRACY, ABSTRACTION, PRINCIPLE…). I will now briefly show how the methodology I advocate here *might* be extended to other troublesome concepts. Dove (2016) recently presented the alleged abstract concept DEMOCRACY as an example of a concept that poses 'a serious challenge' to embodied cognition views. I am not about to provide an argument that embodied cognition is "correct"; instead I want to provide an example of the kind of move anybody interested in concepts can make if they endorse my view.

Does DEMOCRACY look like JUSTICE? In many respects, the answer to this question seems to be 'yes'. We can use the word 'democracy' to talk about all kinds of social hierarchies and complicated relationship structures extending from the international level of organisation, all the way down to the level of members of a family deciding which particular meal to have for supper. According to the general model, DEMOCRACY must be a unitary cognitive resource acquired in such a way that

comports with the theories we have, and that could plausibly play a singular role in mediating behaviours and cognitive processes in situations that extend from the dinner table all the way to the intricate mechanisms by which individuals in certain groups instantiate the governance of nation states on behalf of the people they represent. *And* it would also presumably have to correspond to and pick out something that we would want to call a 'category'. If you are a Fodorian, then you would like to be able to say what its individuation conditions are, and if you are a Barsalou-ian, you would like to be able to say what is part of its network of simulators and what isn't. I am advocating some scepticism towards the *assumption* that this theoretical posit, DEMOCRACY, is going to play a useful role in our theories of cognition. Note that, as it stands, this does not decide the issue either way. I think a consensus about which words of English pick out explanatorily powerful units of a theory of concepts is something that can only be achieved with investigation. So it could be that, after some careful theoretical and empirical work, we decide to include DEMOCRACY in our theory of concepts. However, we should be open to the possibility that the best course of action would be to exclude it. I also think that investigations like these can provide interpretable data and useful debate about the human conceptual system. One of the things I hope to have convinced you of by the end of this thesis is that the concreteness scale probably does neither of these things.

Instead of assuming that there is at least a concept for every word we know, in my view we should step back and think about the specific human behaviours and cognitive processes that we want to investigate, and then work out whether and how a theory of concepts will interact with these investigations. If an alleged concept survives the test I have proposed, then it does work in our theories of concepts, and so it should be thought of as a concept. If it does not survive the test, then a theory of concepts does not have to worry about it. We should posit other mental representations that figure in specific cognitive processes and that mediate specific behaviours, and these other mental representations are more likely to be amenable to analysis no matter which kind of theory you endorse. I suspect that a number of 'abstract' concepts will not survive the test, and ultimately that the concrete-abstract distinction does not amount to anything interesting. Instead, there are just those concepts that we *have* (if you are a realist) or those concepts that work in theories (if you an antirealist).

3.8 Two important objections to giving up the concept JUSTICE

There are (at least) two very important objections to my position that I wish to consider. These two objections are:

1. If some proposed abstract concepts don't actually exist, then how can it be that reliable experimental effects are obtained by measuring responses to 'concrete' stimuli and comparing them to responses to 'abstract' stimuli? Surely this empirical evidence indicates that there must be something neuro-psychologically real and theoretically principled about the concrete-abstract distinction, and that there is a reliable relationship between words and concepts.

2. If there is no unitary cognitive resource corresponds to the word 'justice', then what is the meaning of the word 'justice' – how do we understand each other when we use the word 'justice', or other abstract words for that matter?

I spend the majority of the rest of this thesis answering these two objections and considering issues that my responses raise. In order to respond to objection 1, I first look at a simple statistical summary of a very large concreteness norm database produced recently by Brysbaert et al. (2013). In doing so, I will be able to demonstrate that the concreteness measure has some alarming properties that arguably invalidate it as a psycholinguistic tool. I also show that this problem also applies to similar variables and other databases, such as Connell and Lynott's (2012) modality exclusivity norms, and imageability (Cortese and Fugett, 2004; Schock et al., 2012), which is highly correlated with concreteness. I will also show that concreteness experiments to date have not actually compared responses to concrete stimuli with responses to abstract stimuli. This severely weakens the force of objection 1. In Chapters 5 and 6 I report experiments which were designed to assess the severity of the problems I identify with the concreteness measure. Worryingly, under conditions that should have maximised both the chances of finding a concreteness effect and the magnitude of that effect, these new experiments returned null results in all but one case, in which the concreteness effect was very small. In Chapter 7, I will argue

that evidence for experimental concreteness effects is not actually as strong as it is assumed to be, and therefore that objection 1 is not fatal to the view I am trying to convince you of. In Chapter 8, I will deal with objection 2. My response to objection 2 involves sketching a theory of word meaning that does not identify word meaning with concepts. I show how Relevance Theory (Sperber and Wilson, 1998, 1995), a popular theory of communication, can be adapted so that it does not require that word meanings and concepts are the same things, in such a way as to maintain its considerable explanatory power. I also suggest that in some ways, a non-conceptual account of word meaning is preferable to a conceptual one. In this way, I hope to have provided a compelling response to objection 2.

Chapter 4: Concreteness itself[7]

Recall from Chapter 3 that a pressing objection to my argument relied on the observation that the concreteness measure has produced a huge array of statistically significant experimental effects, which we saw in Chapter 2. The objection was that because we have amassed such a large number of statistically significant concreteness effects, there must be something principled and psychologically relevant about the concrete-abstract distinction. If this were true, then we might conclude that my suggestion that words might not reliably pick out elements of a theory of concepts is wrong, because these effects are all explained by assuming that all words *do* reliably pick out concepts. In this chapter, I start to set out my responses to this objection. I begin by offering some theoretical reasons why we should be cautious about using concreteness scales in psycholinguistic research. Next, I present a statistical analysis of the concreteness norm database produced recently by Brysbaert et al. (2013). I show that that for nearly every single word in the middle of the concreteness scale, the mean concreteness value of a word does not reflect the judgements that individual participants actually made about it. Instead, ratings for words in the middle of the scale are essentially noise. Then, I show that for a great many concreteness experiments reported in the literature, the stimuli in the 'abstract' conditions were not actually abstract. Instead, they were simply those words about which participants disagreed, and for which the concreteness rating is uninterpretable. I also present an argument that the criticisms I raise here should be just as worrying to researchers who prefer large scale regression designs, as opposed to factorial designs (Connell and Lynott, 2012; Kousta et al., 2011). I wish to stress here that my intention is not to single out any of these experiments or researchers for criticism. The analysis that I present was only made possible after the Brysbaert et al. (2013) database was published. My aim is only to draw attention to a problem that I believe has implications for concreteness research.

I spend the rest of this chapter exploring some implications of these observations, and I also consider other psycholinguistic variables that are related to concreteness (Connell and Lynott, 2012; Cortese and Fugett, 2004; Schock et al., 2012). I show that these variables suffer from exactly the same problematic distribution as

[7] This chapter is largely based on the ideas discussed in Pollock (2017)

concreteness. Finally, I show that this problem is not an issue that is general to all subjective rating scales used in psycholinguistics: the Warriner et al. (2013) emotional valence norms do not have the same distribution as the other norms. This shows that the issue is specific to the scales that are used to investigate or operationalise the concrete-abstract distinction.

4.1 The middle of the concreteness scale

We saw in Chapter 2 how concreteness norms are generated. A word's concreteness rating is derived by asking a group of participants, typically numbering between twenty and thirty, to rate that word for concreteness on a Likert scale. A low score indicates that a word is highly 'abstract', whereas a high rating indicates that a word is highly 'concrete'. I will now develop some theoretical concerns about the validity of traditional concreteness norms before turning to a statistical analysis of the Brysbaert et al. (2013) database. Firstly, let us consider the job a participant is being asked to do when she is told to rate a word between, say, 1 and 5 on a scale of concreteness. She is told that 'concrete words are experienced by the senses' (as per the norming instructions introduced in Chapter 2), whereas abstract words are not. As I noted before, this distinction between concreteness and abstractness is widely used, but it is again important to point out that the definition of abstractness is entirely negative. The only description offered of what constitutes abstractness is that it is 'not concreteness'. By crude analogy, someone curious about the difference between solids and liquids would be rightly dissatisfied with the explanation that solidity is a state of matter characterised by structural rigidity, and that liquids are 'not solids', because this definition does not tell this person much at all about the properties of liquids. Attempts to formulate a positive definition of abstractness that identifies its properties are rare. Indeed, as has been pointed out a number of times, it is only recently that researchers have even started to focus on abstract concepts in detail and reverse a historical tendency to focus on concrete concepts exclusively (Crutch and Ridgway, 2012; Hamilton and Coslett, 2008; Wiemer-Hastings and Xu, 2005). This on its own might be reason enough to wonder what the basis is for a participant's decisions when they engage in these rating tasks. But for now I want to focus on the middle section of the concreteness scale, which supposedly contains those words that are of intermediate concreteness.

It is reasonable to assume that for a certain class of words, the interpretation of traditional concreteness norming instructions is relatively straightforward. A participant that is presented with the word 'apple' is likely to have seen, touched,

smelled, and tasted apples throughout the course of their life, and will unproblematically assign apple a high concreteness rating. Similarly, a participant that is presented with the word ‘serendipity’ is likely to reason that since serendipity is a loose association between some coincidental, non-specified events, and is not something that affords direct sensory experience, the word ‘serendipity’ should be assigned a low concreteness rating. However, what are the properties that a word/concept should have in order for it to be assigned a midscale rating? It is difficult to formulate a coherent approach to this task that is predicated on the idea that a given object or idea can be ‘half-seen’ or ‘half-touched’. What does it mean to have intermediate sensory experience of an entity or idea? That is to ask: what is a participant telling us about a word when they rate it a 3 out of 5? They could mean any one of the following:

1) Adding up all of my sensory experience of this object across all five of the sensory modalities, I realise that I have seen and heard it, but never touched, smelled, or tasted it. So I suppose I’ll rate it a 3.

2) One interpretation of this word brings to mind something that cannot be directly experienced, whereas a different interpretation of this word brings to mind something that can be directly experienced. So I suppose I’ll rate it a 3.

3) Sometimes I associate sensory experience with this word, but sometimes I don’t. So I suppose I’ll rate it a 3.

It is certainly possible to imagine more potential approaches. And there is no empirical basis for selecting one of these approaches over another. Furthermore, it is likely that different participants will generate different interpretations for some of the words in any list of words to be normed. Producing examples of this class of words is trivial. When a participant sees the letter-string <deed> presented in isolation, there is no way that a researcher can control for the fact that half the participants may interpret <deed> as referring to a document associated with proof of property ownership (high concreteness value?), and the other half may interpret <deed> as referring to some unspecified action, perhaps involving some element of heroism (low concreteness value?). This problem is not confined to homographs. Metonymies are ubiquitous in English and also introduce ambiguity into the norming process. When asking participants to indicate the concreteness value of the concept that the letter-string <football> refers to, it is impossible to know whether participants interpreted <football> as referring to the ball that is used to play the sport or whether they interpreted <football> as referring to the sport itself. It is likely that some participants

settled on the first interpretation, while some settled on the latter. And it also seems reasonable to allow that different interpretations might be assigned different concreteness ratings. The concept of a spherical ball that can indeed be touched, heard, and seen is likely to be given a higher concreteness rating than the concept of a group endeavour predicated on a system of rules that happens to involve a spherical ball. But these interpretations are not separable in traditional concreteness norming methodologies, and consequently there are a number of words for which it is just not clear what concept it is that the mean concreteness rating is even supposed to reflect.

So far, no one has investigated whether all participants are using the same judgement criteria, or whether they are all using different judgement criteria, or whether they use one judgement criterion for some words and a different judgement criterion for others, or whether different participants are even interpreting words in the same way. This point on its own might be enough to motivate the avoidance of words with a mean value in the middle of a concreteness-abstractness scale. Given that it is not clear what it is that participants are even telling us when they rate a word a 3, we might also wonder how often participants actually use values from the middle of the concreteness scale when making their judgements. As we shall see below, the problems relating to this observation are serious, and midscale words feature frequently in concreteness experiments. One of the reasons that this issue is common is that until recently, the available databases that provide researchers with concreteness rating norms were relatively small. After having to control stimuli for nuisance variables such as word length, frequency, age of acquisition, and semantic category, the pool of items that fits all these constraints is greatly reduced. Recently, Brysbaert et al. (2013) provided a concreteness norm database of 40,000 English words, which dwarfs the previously popular MRC database used in most studies (Coltheart, 1981). This new, larger database allows a statistical analysis of the distributions of concreteness norms across a much larger section of the English lexicon. I now present this analysis and use it to develop the concerns raised in this section.

4.2 A statistical analysis of Brysbaert et al.'s (2013) concreteness database

Brysbaert et al. (2013) collected a new set of concreteness norms for 40,000 English words. Groups of approximately 25 participants rated subsets of the whole list of 40,000 words on a concreteness scale of 1 (very abstract) to 5 (very concrete).

The instructions given to participants were intentionally different to the original Paivio et al. (1968) instructions: 'the instructions stressed that the assessment of word concreteness would be based on experiences involving all senses and motor responses' (Brysbaert et al., 2013, p. 904). This change was made in response to criticisms from Lynott and Connell (2012) that, in producing the original concreteness norms, participants relied too heavily on the visual and haptic modalities when judging how concrete a word was, and neglected auditory, olfactory and gustatory modalities. The full definitions from Brysbaert et al. (2013, p. 906) are reproduced here:

> A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e.g. To explain 'sweet ' you could have someone eat sugar; To explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; To explain 'couch', you could point to a couch or show a picture of a couch).
>
> An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning).

The other main difference between the Brysbaert et al. (2013) norms and the MRC database norms is that whereas the MRC database norms are based on a seven-point scale (1 = abstract, 7 = concrete), the Brysbaert et al. (2013) norms use a five-point scale. This change was introduced on the basis of Laming's (2003) argument that humans are incapable of reliably making use of more than five categories in judgment tasks. Instead of providing increased resolution, a scale with more than five points really just adds unnecessary noise to the rating process. As an aside, it is worth pointing out here that the conclusions drawn by Laming (2003) are actually somewhat stronger and prompt further concerns surrounding the instructions given to participants in these norming tasks. Laming claims that as well as not being able to consistently distinguish between more than five categories, humans are also much

less able to make absolute judgements than they are to make relative judgements. In this context, the implication would be that participants in concreteness norming studies are only able to consistently judge to what degree a presented word is concrete or abstract *in comparison to a reference word.* In the instructions given in Paivio et al. (1968) and Brysbaert et al. (2013), participants are given exemplars of the extreme ends of the concrete-abstract continuum. In the Paivio instructions, the word *chair* is given as a paradigmatic concrete item, and the word *independence* is given as a paradigmatic abstract item. In Brysbaert et al. (2013), three examples of highly concrete items are given (*sweet* (adjective), *jump*, *couch*) and one example of a highly abstract item is given (*justice*). But in neither set of instructions is an example of an intermediately concrete item given. So not only is there variability in how participants interpret instructions concerning intermediately concrete items, but participants are also not provided with a reference intermediate item that indicates what might constitute intermediate concreteness. Laming shows that in many circumstances, category judgments made without a reference point are essentially random. One solution might be to simply provide participants with some examples of paradigmatic intermediately concrete items. But in doing this, the researcher is effectively imposing their own interpretation of what constitutes 'intermediate concreteness', and as was demonstrated in the previous section, there is no basis for choosing one set of criteria for intermediate concreteness over another set. This is a further problem with traditional concreteness norms.

The mean value of a group of participants' judgments about the concreteness of a stimulus word is assumed throughout the literature on word concreteness to be a useful approximation of that word's position on a hypothesised concrete-abstract continuum. I shall now demonstrate that this assumption is false. Consider the following example datasets, where each line represents a word rated by four participants in a concreteness norming study, and each individual number in square brackets represents an individual participant's judgment about the concreteness of that word on a scale of 1 to 5:

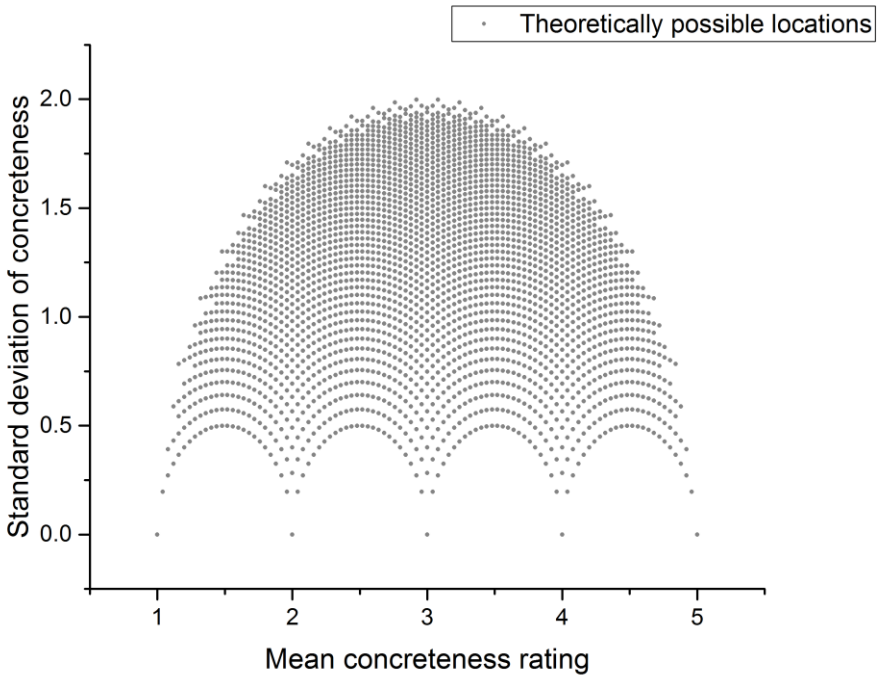
[1 1 1 1] Mean = 1, Standard Deviation = 0 Highly Abstract
[5 5 5 5] Mean = 5, Standard Deviation = 0 Highly Concrete
[3 3 3 3] Mean = 3, Standard Deviation = 0 Intermediately Concrete
[1 5 1 5] Mean = 3, Standard Deviation = 2 Intermediately Concrete

The standard deviation of a dataset is a measure of the average distance between all the data points in that dataset and the mean value of all data points in the dataset. In norming tasks, the standard deviation of a set of ratings is therefore a blunt index of

the extent to which participants agreed with each other about how a word should be rated. To make this point clear, consider the first three words in this example dataset. Participants all unanimously agreed with each other about the concreteness of these words, and therefore the standard deviation of the ratings of these words is 0, and the mean rating accurately reflects the participants' individual judgments. If this hypothetical situation occurs in actual concreteness norms, then we can have some hope that the concerns outlined in the previous section are unfounded: unanimity of judgment might suggest at least some degree of unanimity of interpretation. However, now consider the fourth word in the example dataset. Here, the mean value is 3 despite the fact that not a single participant judged that word to be intermediately concrete. In fact, two participants judged this word to be as abstract as possible and assigned it a 1, whereas the other two participants judged this word to be as concrete as possible and assigned it a 5. Half the participants totally disagreed with the other half, and the standard deviation is 2: on average, all participants' judgments are 2 scale positions away from the mean value of the dataset. This would be a very troubling situation indeed, because it shows that the mean concreteness value of this fourth word bears little relation to participants' responses. These participants were either interpreting concreteness norming instructions differently, disagreeing with each other about what concreteness is, or were supplying ratings on the basis of differing interpretations of the target word. Indeed, a combination of all three factors is possible.

Recall that in the Brysbaert et al. (2013) norms, 25 participants rated each word on a scale between 1 and 5. If a dataset contains 25 numbers (in our case, 25 individual concreteness judgments), all of which are integers between 1 and 5, then there are a finite number of possible combinations of means and standard deviations for that dataset. Figure 4-1 below plots all of these possible combinations:

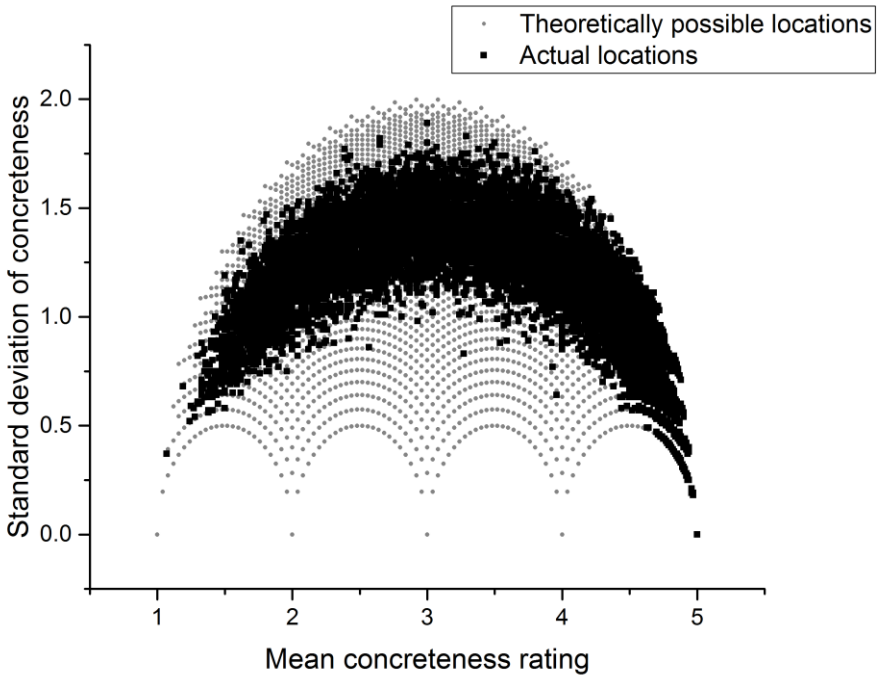
Figure 4-1 Theoretically possible means and standard deviations for concreteness ratings in Brysbaert et al. (2013)



Note how, at the extreme ends of the x-axis, only a standard deviation of 0 is possible because for a mean value to be 1 or 5, all 25 participants must have rated an item as 1 or 5, respectively. However, in the middle of the scale, the disagreement that is theoretically possible increases, reaching a peak at mean value ~3, standard deviation 2. This is exactly analogous to the fourth word in the example dataset introduced above, except that in this case the dataset would consist of 12/13 ratings of 1, and 12/13 ratings of 5, instead of two ratings of each value. Crucially, it is still theoretically possible for a data point to occur with a mean value located in the middle of the scale, but with a relatively low standard deviation. To relate this point to concerns surrounding concreteness norms, it is still clearly theoretically possible for participants to more or less consistently agree that a word is of intermediate concreteness.

Now, consider Figure 4-2 below, which plots the *actual* mean concreteness value and standard deviation of every noun in the Brysbaert et al. (2013) concreteness norm dataset (n = 14,592) over the top of the *theoretically possible* combinations depicted in Figure 4-2:

Figure 4-2 Actual means and standard deviations for concreteness ratings in Brysbaert et al. (2013)



The pattern is striking. At the extreme concrete end of the scale, there are many items with high concreteness ratings and relatively low standard deviations, indicating that participants more or less agreed in their judgments about how to rate these words. At the extreme abstract end of the scale, there are likewise words with low concreteness ratings and relatively low standard deviations, although not to the same extent as at the extreme concrete end. However, in the middle of the scale, there is an obvious rise in standard deviation. There are only a handful of words with a mean value near 3 and a standard deviation even slightly below 1. Indeed, there is a large class of words with a standard deviation well over 1 ranging from mean values of 1.5 to 4.5. What does this indicate? It indicates that for a great number of items, participants were not agreeing in their judgments of how concrete a stimulus word was. At mean values of 2 and 4, there are many cases of standard deviations above 1. Ratings on this scale can only take integer values between 1 and 5. This means that for many of the words with a mean value of 2 or 4, some participants must have been judging these words as belonging to opposite ends of the concreteness scale to the position that the mean value suggests that word belongs to. This phenomenon is problematic for the assumption that concreteness should be treated as a continuous variable. This is because in a vast number of cases, participants' judgments tended not to be continuous. They tended to be binary. Participants were using values of 1, 2, 4 and 5

in producing these concreteness norms, and tended to avoid using 3. Furthermore, in many cases, some participants were judging a word as a 1 ('totally abstract'), whereas others were judging that same word as a 4 ('somewhat concrete').

Given these methodological issues, it might be considered somewhat surprising that concreteness effects are so widely reported in the psycholinguistic literature. If a large section of the hypothesised concreteness spectrum is actually a procedural artefact resulting from erroneously assuming that a mean value necessarily reflects the individual data points from which it is derived, then it is unclear what phenomenon it is that concreteness effects are actually indexing. One potential explanation immediately presents itself: generally, when investigating the effect of a variable, researchers try to choose stimuli that maximise a change in this variable in order to generate the maximum possible effect. It is therefore possible that empirical concreteness research might not suffer too badly from the problem of binary disagreements concerning midscale items, because researchers will have aimed to pick stimuli from the extreme ends of the scale, and these polar items are less subject to disagreement. However, if it turns out that a significant number of concreteness studies include stimuli that suffer from the disagreement phenomenon, then this poses an explanatory problem concerning evidence in favour of processing differences between abstract and concrete items. The typical finding is that there are processing advantages for concrete items relative to abstract items, and the typical explanation of this finding is that concrete items and abstract items have different neurologically instantiated formats and/or structural relationships. If a significant number of stimuli included in an abstract or concrete experimental condition actually come from the middle of the concreteness scale, then the typical claim that there are processing differences between concrete items and abstract items is no longer supported by the data. This is because for those words with high standard deviations, half of the participants who produced the concreteness measure for that word judged it to be abstract, and the other half judged it to be concrete. Therefore, there are no grounds for calling these words 'concrete' or 'abstract' in the first place.

4.3 Stimuli featured in concreteness experiments

I now present a survey of the stimuli used in empirical concreteness studies published over the last thirty years. The mean concreteness rating and the standard deviation of the concrete and abstract stimuli in these experiments are plotted against the nominal subset of the Brysbaert et al. (2013) concreteness norms. This provides

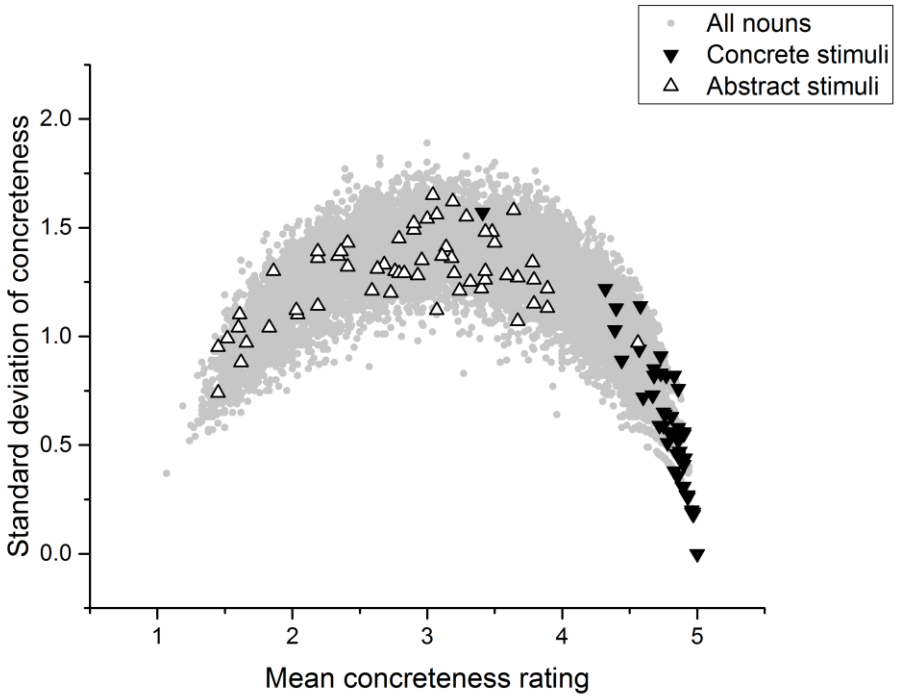
us with a graphical representation of the extent to which stimuli in concreteness experiments avoid the disagreement problem, or fall foul of it. Again, I stress that these studies have not been singled out for criticism. My aim was to choose studies that have at least one of two properties. They have either been frequently cited in the concreteness literature and therefore contribute significantly to background assumptions made in concreteness studies and the discipline of psychology in general, or they attempt to resolve a specific theoretical dispute and therefore make substantive claims on the back of the empirical data they report. A representative sample of experimental paradigms is also included. A further, more pragmatic reason for choosing the following studies over others is that their stimuli sets were, laudably, made available for scrutiny. The stimuli of the following studies will be analysed in reverse chronological order:

Table 4-1 - Studies included in stimuli analysis

Study	**Paradigm**
Kroll and Merves (1985)	Lexical decision
De Groot (1989)	Word association
Binder et al. (2005)	Lexical decision
Romani et al. (2008)	Word recall

Romani et al. (2008) report a series of experiments designed to investigate whether concreteness has an effect on how well participants can recall items held in Short Term Memory (STM). Participants were read lists of words, and then asked to immediately write down the words that they had heard either in the order that they had heard them, or in an unordered 'anything goes' format. Lists of words were supposed to consist entirely of concrete items, or entirely of abstract items. Over the four experiments reported, there were a variety of manipulations in order to investigate fine-grained distinctions between various hypotheses generated by different models of STM, but in general, Romani et al. (2008, p. 312) report 'consistent, positive effects of concreteness in tasks tapping immediate recall of items in the proper order (serial recall), recall of items independent of order (free recall), and recall of item positions (matching span and order reconstruction)'. Now consider Figure 4-3 below, which plots the stimuli used in these experiments against the nominal section of the Brysbaert et al. (2013) norms:

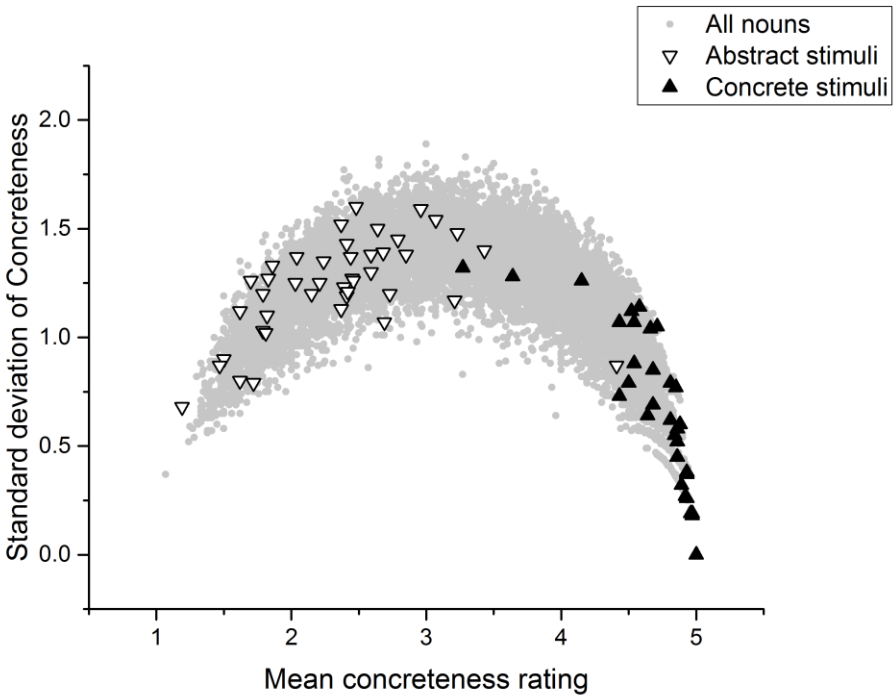
Figure 4-3 Stimuli featured in Romani et al. (2008)



Again, and unfortunately, the pattern is striking. The concrete stimuli, depicted as black triangles, are where they 'should' be: concentrated at the extreme end of the concreteness scale and with relatively low standard deviations. These items are unequivocally concrete. However, this is clearly not the case for the allegedly abstract stimuli depicted as white triangles. The abstract stimuli have actually been taken from across the entire range of the scale, and a great many of them have standard deviations well above 1. This presents a problem for the assumption that the phenomenon under discussion (the effects of a psycholinguistic variable on STM processing) necessarily has anything to do with concreteness. Instead of making a comparison between concrete and abstract items, the comparison is actually between concrete items on the one hand, and a heterogeneous group of items on the other, a large subset of which have mean ratings that do not reflect individual concreteness judgments. Indeed, in order for the standard deviations of these items to be as high as they are (greater than 1), some participants must have been judging them to be relatively concrete. The stimuli that suffer from high disagreement (that is, those stimuli with high standard deviations) therefore cannot be said to be unequivocally 'abstract' in the same way that the concrete items are unequivocally concrete. It is difficult to argue that, despite these problems, the concreteness effect still *necessarily* obtains, because the comparison here is simply not between concrete items and abstract items.

Binder et al. (2005) report a lexical decision study in which participants were presented with single letter-string stimuli one item at a time, and asked to indicate as quickly as possible via a button press whether each letter-string represented a real English word, or non-word. Binder et al. (2005) also acquired fMRI data from participants performing this task in an attempt to localise distinct concrete and abstract brain activation areas. There were three experimental conditions: concrete words, abstract words, and non-words. Binder et al. (2005, p. 907) report 'a reliable advantage of concrete over abstract items' in both the reaction time data and error rate data. Concrete items were, therefore, processed faster and more accurately than abstract items. Now consider Figure 4-4, which plots the stimuli used in this experiment against the entire distribution of the nominal section of the Brysbaert et al. (2013) norms:

Figure 4-4 Stimuli featured in Binder et al. (2005)

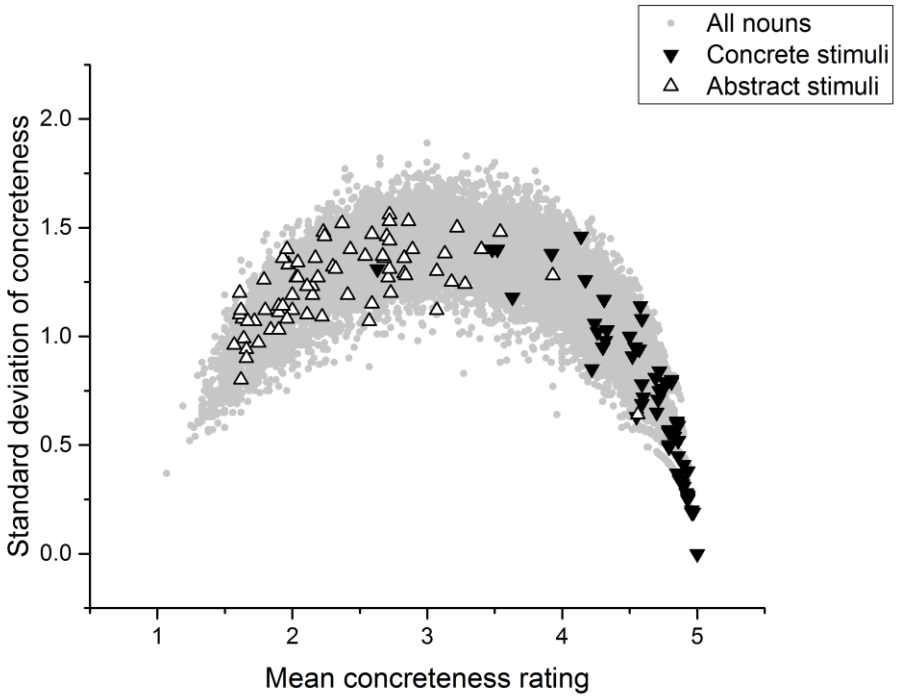


Although the pattern is not as pronounced as that seen in the distribution of the stimuli used in Romani et al. (2008), it is similar and still concerning. Once again, the concrete stimuli are relatively unproblematic, and the vast majority of them are located at the extreme end of the scale and have relatively low standard deviations. However, the abstract stimuli show a far greater range, and their standard deviations (M = 1.23) are still high compared to the standard deviations of the concrete stimuli (M = 0.52). A simple independent samples t-test reveals that this difference is statistically

significant ($t = 10.132$, $df = 88$, $p < 0.001$). The abstract stimuli featured in Binder et al. (2005) are therefore subject to the same criticisms as the stimuli featured in Romani et al. (2008). The mean ratings of a large subset of the abstract stimuli simply do not reflect participants' judgments about those stimuli, and as a consequence it is not clear that they should be thought of as abstract. Once again, the comparison being made is not actually between concrete and abstract items, but rather between items that norming participants unanimously agree on as being concrete, and items that norming participants disagree about how to rate. It is therefore not safe to draw conclusions about *concreteness* effects on the basis of these stimuli, because abstractness has been confounded with 'disagreement about concreteness'.

De Groot (1989) reports a series of experiments designed to investigate potential interactions between the frequency with which a word appears in written discourse, the 'imageability' of the referent of that word, and the ease with which participants make word associations in response to it. Roughly, the imageability of a word is the ease with which one can generate a mental image of the referent of that word. Note that de Groot uses the terms 'imageability' and 'concreteness' more or less interchangeably. This practice is common in research on concreteness (and, indeed, imageability) because the two variables are highly correlated, despite the fact that they were originally conceived of as separate constructs (Paivio et al., 1968). In any case, de Groot (1989) reports that on all measures, concrete (i.e., highly imageable) items were responded to faster and more consistently than abstract (i.e., less imageable) items. Although the strategy of using imageability norms but then analysing results in terms of concreteness theories potentially adds to the confusion surrounding the reliability of concreteness effects rather than decreasing it, the Brysbaert et al. (2013) norms still provide a measure of the distribution of de Groot's (1989) stimuli across the concreteness scale. Furthermore, as we shall see in later sections of this chapter, imageability as a variable suffers from much the same problems as concreteness. Consider Figure 4-5, the format of which should by now be familiar:

Figure 4-5 Stimuli featured in de Groot (1989)

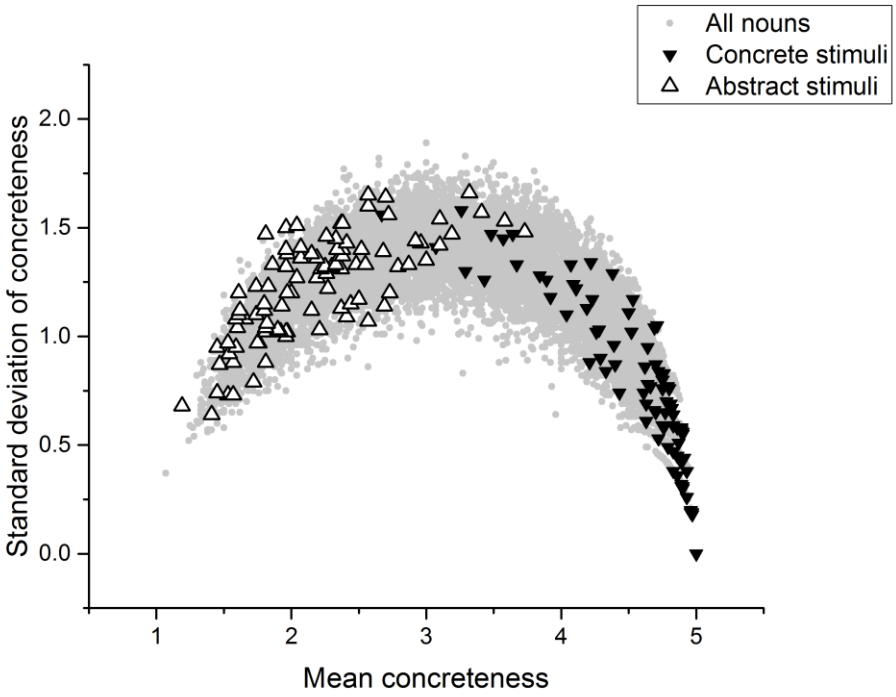


Once again, concrete stimuli tend to have relatively low standard deviations (M = 0.56), whereas the abstract stimuli are evenly distributed across a larger section of the scale, and tend to have high standard deviations (M = 1.24). An independent samples t-test reveals that this difference in standard deviation is statistically significant ($t = -12.186$, $df = 143$, $p < 0.001$). Therefore, as with the two previous studies, there is a confound between the abstractness of the stimuli in the 'abstract' condition, and the extent to which participants were able to consistently judge the concreteness of those stimuli in the first place.

Finally, Kroll and Merves (1985) report a series of experiments designed to isolate the precise conditions under which concreteness effects appear in Lexical Decision. They note that previous research on concreteness effects in Lexical Decision had provided conflicting results: sometimes the concreteness effect obtained, and sometimes it did not. They examine the predictions of Dual Coding Theory about the outcomes of stimuli presentation manipulations. Stimuli were either presented in blocks by word type, or presented in randomly mixed lists containing both abstract and concrete words. Kroll and Merves (1985) report small (and on some measures absent) concreteness effects in some experiments, but identify a large concreteness effect in an experiment in which blocked concrete items were presented

before blocked abstract items. Figure 4-6 below plots the stimuli featured in these experiments:

Figure 4-6 Stimuli featured in Kroll and Merves (1985)



At first glance, the distribution of the stimuli across conditions in Kroll and Merves (1985) might appear to be better than the distributions of the previous studies. The sample of concrete stimuli is similar to that of the previous three studies. However, there are a greater number of abstract stimuli drawn from the extreme end of the scale and their standard deviations seem lower. However, an independent samples t-test reveals that the difference in standard deviations between the abstract (M = 1.22) and concrete stimuli (M = 0.74) is still statistically significant (t = -10.285, df = 189, $p < 0.001$).

Crucially, these four experimental reports are not special cases when it comes to the properties of 'abstract' stimuli. Table 4-2 below presents a number of experimental concreteness studies from a wide variety of paradigms, and a summary of the concreteness values and standard deviations of the stimuli featured in their experiments. The abstract-midscale stimuli pattern applies to every single experiment.

Table 4-2 Concreteness statistics in various experimental paradigms

Article	Type of Data	Experimental paradigm	Concrete		Abstract	
			Mean	SD	Mean	SD
Kroll and Merves (1985)	Behavioural	Lexical Decision	4.55	0.74	2.17	1.22
de Groot (1989)	Behavioural	Word Assoc.	4.66	0.6	2.36	1.24
Paivio et al. (1994)	Behavioural	Recall	4.83	0.47	2.29	1.28
Gee et al. (1999)	Behavioural	Recall	4.73	0.57	3	1.33
Binder et al. (2005)	fMRI	Lexical Decision	4.76	0.52	2.34	1.23
Crutch and Warrington (2005)	Patient	Word matching	4.83	0.46	3.53	1.18
Sabsevitz et al. (2005)	fMRI	Sem. judgement	4.86	0.45	2.58	1.31
ter Doest and Semin (2005)	Behavioural	Recall	4.72	0.57	2.45	1.26
Lee and Federmeier (2008)	EEG	Sem. judgement	4.41	0.88	2.27	1.24
Huang et al. (2010)	EEG	Sem. judgement	3.82	1.17	2.53	1.21
Skipper-Kallal et al. (2015)	fMRI	Deep thought	4.44	0.81	2.38	1.22
Jager and Cleland (2016)	Behavioural	Lexical Decision	4.62	0.64	3.29	1.19

These studies were chosen simply because they reflect a range of experimental paradigms (lexical decision, recall, semantic judgement, word association, picture-word matching), data types (behavioural, fMRI, EEG), and include both neurotypical and patient populations. They also included their stimuli sets in their experimental reports, although it is important to note that for Sabsevitz et al. (2005) and Lee and Federmeier (2008), only a sample of the stimuli were available. For every study but one listed in this table, the mean standard deviation of the stimuli in the concrete conditions was below 1, while the mean standard deviation of the stimuli in the abstract conditions was above 1. The only exception is Huang et al. (2010), in which the standard deviations for both stimuli sets were relatively high. Looking at the distributions displayed above in the figures above, it is clear that the only way these statistics could be obtained is if the midscale disagreement problem applied to all of the abstract stimuli sets of the experiments depicted in table 4-2.

This brief survey of empirical concreteness research undertaken over a period of three decades shows that no study has avoided confounding the mean concreteness ratings of their stimuli with the variability in those ratings (as indicated by the standard deviation) to one degree or another. The theoretical problems arising from the variability of interpretation inherent in both single-word presentation paradigms and concreteness norming instructions indicate that the middle of the concreteness scale is unlikely to help us develop an applicable model of the human conceptual system. This, in combination with the finding that the 'abstract' stimuli in concreteness experiments are in many cases not actually abstract, but are instead exactly those items that are methodologically problematic, raises an obvious question: how and why is it that concreteness effects appear to be so robust despite these methodological issues?

One answer to this question is simply to point out that, actually, concreteness effects are not that robust, and the literature is marked by conflicting findings. In Chapter 2, we saw that lexical decision experiments have not consistently produced concreteness effects, and in some cases have even produced abstractness effects. We also saw that the fMRI data is highly variable, and that the very same areas of the brain have been argued to be especially essential to the processing of each word type over the other. One potential explanation of this variability is that different experimental tasks and stimulus lists will necessarily introduce some variability into the data produced by different experiments. However, on top of this, we now have evidence that the stimuli that were used in these experiments were simply not good representatives of the alleged category distinction between concrete and abstract words. Concrete stimuli are generally unproblematic, but abstract stimuli are simply those words about which participants tended to disagree when the concreteness measure was generated. We might expect that responses to these words would be inherently more variable in other respects than those responses to words about which participants tend to agree. With this additional source of noise in experimental designs, perhaps it is no wonder that lexical decision and fMRI paradigms have produced such variable results.

However, it is still true that in list memory (Allen and Hulme, 2006; Miller and Roodenrys, 2009; Romani et al., 2008; Walker and Hulme, 1999) and EEG paradigms (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000), consistent concreteness effects have been obtained. Participants are better at recalling concrete words than abstract words in list memory paradigms, and N400 amplitudes to concrete words are higher than N400 amplitudes to abstract words. The purpose of Chapters 5 and 6 is to report new experiments that controlled

for the problems outlined in this chapter, in order to determine what happens when the concrete-abstract contrast is maximised, and concrete stimuli are compared to 'truly' abstract stimuli. Because list memory and EEG paradigms seem to produce consistent effects even though they also featured problematic stimuli, these are the paradigms we shall focus on.

4.4 Concreteness and multiple linear regression

All of the experiments mentioned above are factorial designs in which high concreteness stimuli were compared to low concreteness stimuli. The mid-scale disagreement phenomenon raises problems for this kind of study, because the stimuli featured in the abstract conditions were not unequivocally abstract. Therefore, even if we do obtain statistically significant differences between concrete and abstract conditions, it isn't clear that this difference can be attributed to the concreteness variable, because it isn't clear that the variable does indeed separate two distinct classes of stimuli. However, you might suppose that large scale multiple linear regression (MLR) analyses are immune to these criticisms (Kousta et al., 2011; Lynott and Connell, 2012). You might suppose this because these MLR analyses don't compare concrete stimuli with abstract stimuli explicitly; they quantify the variance in some dependent measure that is explained when concreteness is added to a linear regression model. If a statistically significant portion of variance is explained this way, then surely it's okay to use the concreteness variable in this kind of experiment. I will now argue that, unfortunately, this is incorrect for two reasons.

The first problem facing MLR approaches is that they assume that concreteness is actually a linear variable: it is a requirement of MLR that variables are such. However, the pattern of participants' judgements reported in section 4.2 shows that concreteness is not a linear variable. Participants do not use the whole scale, and tend to pick values from the extreme ends of the scale. Participants treat concreteness as a binary variable; not a linear variable. This on its own invalidates models that assume that concreteness is a linear variable (the p-values produced by these models are arguably uninterpretable). The second issue is that, as we have already seen, there are conflicting findings when it comes to the effect of concreteness on behavioural measures, even in large scale regression analyses. Connell & Lynott (2012) find that concreteness is negatively correlated with decision latency. Kousta et al. (2011) find that concreteness is positively correlated with decision latency. Brysbaert et al. (2016) do not find a statistically significant effect of concreteness on decision latency either way. In my view, this is exactly what we would

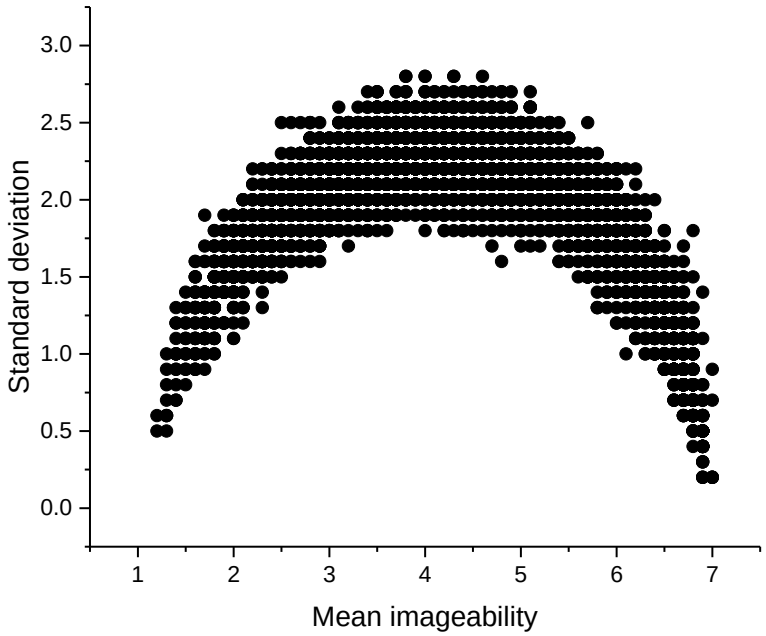
expect if the assumptions of these models were inappropriate, and/or if the models contain lots of noise. Ultimately, the midscale disagreement phenomenon reduces our confidence in the interpretations of large scale regression studies just as much as it does for factorial designs.

4.5 Other subjective sensorimotor rating scales

Before we move on to Chapters 5 and 6, I want to discuss a potential issue with the analyses I have just presented. You might accept that there is a problem with the Brysbaert et al. (2013) concreteness database, and that, therefore, we should be careful when we select stimuli from it. However, there are other concreteness databases (Coltheart, 1981), and other variables that measure very similar things to concreteness, such as imageability (Cortese and Fugett, 2004; Schock et al., 2012), and modality exclusivity norms (Lynott and Connell, 2012). Even accepting the problem with the Brysbaert et al. database, we shouldn't assume that these other norming databases have the same problem. If these other measures contain information that is interpretable right across the scale, and with relatively low variability in ratings, then perhaps the worries I have raised in this chapter aren't so pressing after all. We can simply use these other databases instead of the Brysbaert et al. database, and the concreteness measure is still valid in the general case. I now want to show that, unfortunately, the midscale variability problem applies to all of these databases as well: it is general to all subjective rating scales that are designed to measure the extent to which participants associate modal (sensory) information with words. Perhaps this is not too surprising because they are derived in much the same way as concreteness (by taking the mean value of a set of individual judgements about depth of sensorimotor experience).

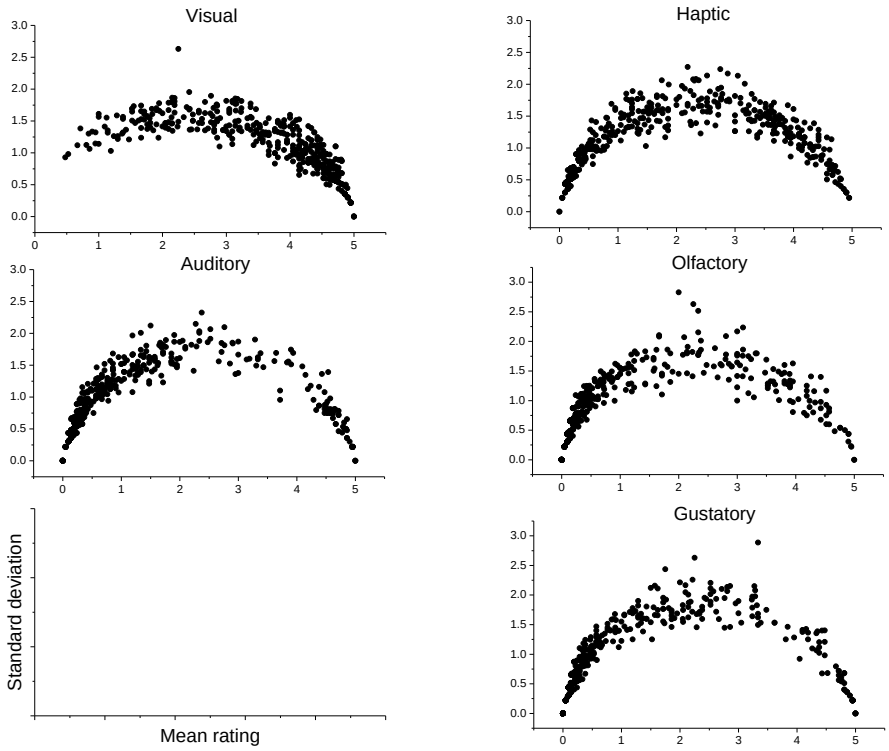
Figure 4-7 presents a mean/standard deviation plot of the imageability ratings of 6,000 words amalgamated from two databases (Cortese and Fugett, 2004; Schock et al., 2012). Imageability is a measure of how easy it is to generate a mental image of the referent of a word. It is so highly correlated with concreteness that the two variables have often been used interchangeably in the literature (as we saw with de Groot (1989)).

Figure 4-7 Means and standard deviations of imageability ratings for 6,000 words (Cortese and Fugett, 2004; Schock et al., 2012)



The distribution is identical to that of the concreteness measure. A similar pattern emerges in Lynott and Connell's (2012) Modality Exclusivity Norm (MEN). MEN is essentially measuring the same thing as concreteness, but it provides more information because it features ratings for all 5 primary sensory modalities (sight, sound, touch, taste, smell). A low rating indicates that the referent of a word offers little experience in a given modality; a high rating indicates that a referent offers a lot of experience. Each word is rated on all five modalities. This results in a five element vector from which various measures can be derived (mean sensory experience, maximum sensory experience, Euclidean distance from origin, and so on). Figure 4-8 displays mean/standard deviation plots of all 400 words in the MEN for the five sensory modalities.

Figure 4-8 Means and standard deviations of Lynott and Connell's (2012) Modality Exclusivity Norms



What is striking here is that even with just 400 words, the familiar shape of the distribution is clearly apparent. I do not think that we can ignore the fact that all of these datasets have the same problematic distribution. It is likely to be a result of the question that we ask participants when we generate these measures. When we present depth of sensorimotor experience as a scale, we are implicitly committing to the idea that is possible for an entity to be 'half-real', or 'half in space-time', or 'half-seeable'. The distributions of these semantic variables tell us that participants tend to reject this idea: they do not use midscale values.

One solution might be to specify explicitly what we want the middle of these scales to represent, and to provide examples of midscale words for participants so they have something to anchor their judgements to. Whether something along these lines would usefully decrease variability in the middle of the scale is an open question, but a potential issue here is that it is very difficult (for me) to think of a construct that could serve as a midscale anchor between 'concreteness' and 'abstractness'. Even more worryingly, there are relatively few words in the abstract half of the scale with low standard deviations, and I think this suggests that the concrete/abstract dichotomy is just not well formed. Typically, concreteness research has focused on nouns rather than adjectives or verbs. Even starting with a set of 40,000 words, the

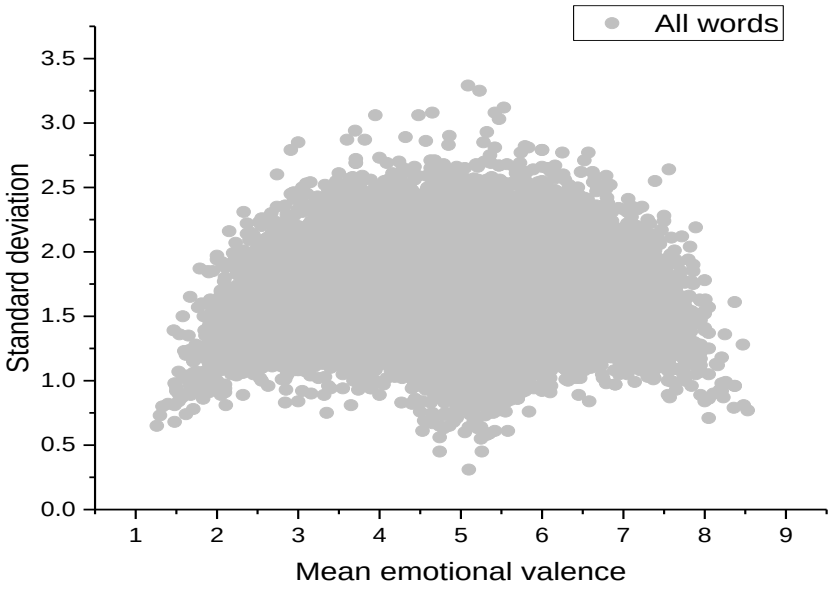
number of nouns in the Brysbaert et al. (2013) norms that have the following properties:

A mean rating of 2 or below (are highly abstract)
A standard deviation of 1 or below
Were known by 100% of the norming population

is only 275. Of these, a small but non-trivial number are either idiomatic fragments (‘amuck’) or morphologically complex rarities (‘purposefulness’) that we might be reluctant to include in stimulus lists. In contrast, there are 2,888 well-known nouns with mean ratings of 4 and above, and standard deviations below 1. I think this fact should motivate caution concerning the utility of the concreteness measure. Ultimately, the measure is supposed to tap into a fundamental, neuropsychologically real distinction between different kinds of concept. It is worrying that the rating is only interpretable for a small number of nominal ‘concepts’ at the abstract pole.

The final thing I want to consider in this chapter is emotional valence norms (Warriner et al., 2013). I have shown that there is a serious issue with all subjective ratings measures that involve sensorimotor associations to words. Participants don’t use the middle values of these scales, and experiments have tended to feature those words for which the measure is uninterpretable. We might worry even further that this problem applies to *all* subjective rating scales hypothesised to measure something about the ‘semantic’ system. Happily, we can rest assured that this isn’t the case. Figure 4-9 below plots the means and standard deviations of the emotional valence ratings of words gathered by Warriner et al. (2013) (n = 13,900). Warriner et al. do present this plot and touch on this issue, but they do not raise exactly the same point as the one I want to focus on here.

Figure 4-9 Means and standard deviations of Warriner et al.'s (2013) emotional valence norms



A score of 1 indicates extremely negative emotional valence, 5 indicates neutrality, and 9 indicates extremely positive emotional valence. Looking at figure 10, it should be possible to select unequivocally negative, neutral, and positive words for use in experiments: there are some words at mean ratings 1, 5, and 9 with low standard deviations. This is obviously a good thing. It also shows that the midscale disagreement phenomenon specifically applies to the concreteness measure and other measures based on more or less the same idea, and so it is concreteness experiments specifically that the issues I have considered here affect. However, I should note that because the middle of the emotional valance scale is a neutral point between two extremes, words with high standard deviations are especially problematic. This is because a 5 is supposed to indicate emotional neutrality. But if a word has a mean of 4-6 but a standard deviation of 2 or more, that means that, on average, participants actually associate moderate to large emotional responses with that word. Some participants associate positive emotions with the word, but others associate negative emotions with it. There are quite a few words that *look* neutral, but in fact are not. Some examples are:

Cell	Mean: 4.09	SD: 2.69
Sushi	Mean: 6.25	SD: 2.77
Gym	Mean: 5.84	SD: 2.52

Similarly, if a word has a mean emotional valence of, say, 3, but a standard deviation above 1.5, that means that some people report a very strong negative response to that word whereas some people report little or no emotional response at all. So if a researcher is interested in comparing responses to neutral words with responses to emotionally valenced words, they should definitely avoid words with high standard deviations for emotional valence, because they will add a significant amount of noise to the experimental design. One positive thing to note is that for the emotional valence measure, a high standard deviation is potentially problematic but it is still *interpretable.* It makes sense that different people will associate different emotions with certain words. It also makes sense to think of our emotional responses as graded. I think this is a key difference between the sensorimotor experience variables and the emotional valence measure.

4.6 Summary of Chapter 4

To recap: in this chapter I have started to set out my response to objection 1. Objection 1 was that the concreteness measure has produced a huge array of statistically significant experimental effects. Because we have amassed such a large number of statistically significant concreteness effects, it might seem like there must be something principled and psychologically relevant about the concrete-abstract distinction. However, in this chapter I hope to have shown two things. Firstly, the measure from which these experimental effects were ultimately derived has seriously worrying statistical properties that, arguably, invalidate it is as a psycholinguistic tool. For almost any word in the middle of the Brysbaert et al. (2013) scale, the mean value of the participants' concreteness ratings does not, in fact, reflect the judgements that they made. We therefore have no basis for calling these words concrete or abstract: the measure is uninterpretable because participants disagree about how to apply it. Furthermore, concreteness is often modelled as a continuous variable in both theory and in linear regression models (Brysbaert et al., 2016; Connell and Lynott, 2012; Kousta et al., 2011), but participants don't behave as if it is actually a continuous variable. In a survey of concreteness experiments conducted over a period of three decades and employing a variety of paradigms and dependent variables, we found that every one of these experiments featured stimuli that were not unequivocally abstract. Instead, those 'abstract' stimuli tended to come from the problematic middle section of the concreteness scale. Therefore, there is reason to doubt that *concreteness* effects have actually been demonstrated, because up until this point

we have not typically been comparing concrete stimuli with abstract stimuli. I suggested that using the problematic midscale stimuli in place of truly abstract stimuli may have introduced a large source of unwanted noise into experimental designs, and this might partially (I stress partially) explain why in some paradigms, such as lexical decision and fMRI, concreteness experiments have produced conflicting and inconsistent results. I showed that the same midscale disagreement problem arises in other databases constructed from measures that are very similar to concreteness, such as imageability and Modality Exclusivity Norms (Cortese and Fugett, 2004; Lynott and Connell, 2012; Schock et al., 2012). This indicates that the midscale disagreement phenomenon is not specific to Brysbaert et al.'s database, but that it is a problem with subjective rating scales of sensorimotor experience in general. Finally, we saw that emotional valence rating scales (Warriner et al., 2013) don't suffer from this problem, so it really is just an issue with these sensorimotor scales, and, therefore, a problem specific to theories and experiments that make use of the concreteness construct.

I think these considerations seriously undermine the claim that concreteness effects are 'reliable', because it is not clear that we have actually implemented the concrete-abstract distinction experimentally. However, I have not presented evidence *against* the possibility that there are such things as concreteness in psycholinguistic experiments. And, although there are certainly problems with some of the 'abstract' stimuli featured in concreteness experiments, it is still true that a statistically significant difference between concrete and 'abstract' conditions has been obtained in many paradigms and by many different teams of researchers. In the next two chapters, I report experiments that were designed to investigate what happens if we are careful to maximise the experimental contrast between concrete and abstract stimuli, and only choose words for which the concreteness measure is interpretable when making this contrast. Note that the expectation here, if there are such things as concreteness effects, is that under these conditions the magnitude of any concreteness effect should be maximised, and the chances of obtaining them should also be maximised. This is because we will have increased the magnitude of the experimental contrast by increasing the distance between concrete and abstract stimuli, and only choosing those stimuli that 'truly' belong in those conditions. In Chapter 5, I report three list memory experiments. In Chapter 6, I report an EEG sentence processing experiment. I chose list memory and EEG because these seem to be the two kinds of paradigm that produced the most consistent concreteness effects, as we saw in Chapter 2. Paradoxically, in two out of three list memory

experiments, marginal evidence in favour of the null hypothesis of no effect was obtained. In the third list memory experiment, a small concreteness effect was obtained. In the EEG experiment reported in Chapter 6, reasonably strong evidence in favour of the null hypothesis of no effect was obtained. In Chapter 7, I draw these results together and summarise what I take them to show: evidence for concreteness effects is not very strong after all. In this way, I hope to have provided a response to objection 1.

Chapter 5: Concreteness effects in list memory experiments

In this chapter I will focus solely on the question of whether concreteness effects obtain in a very specific type of list memory task. First, I shall briefly introduce the central issue at hand, namely whether default concreteness effects occur in list memory experiments. Next, I present a survey of list memory experiments that investigate concreteness effects conducted over the last 40 years. I argue that every one of these experiments is flawed in various ways, and that these flaws might *potentially* contribute to an alternative explanation of why default concreteness effects are prevalent in the list memory literature. I then report three new experiments that were designed to avoid these flaws. Two of these experiments produced marginal evidence in favour of the null hypothesis. The final experiment produced a small concreteness effect. I discuss the new results presented here in relation to the results of previous list memory experiments and consider the theoretical issues that pertain to these results. Finally, I conclude that the evidence for default concreteness effects in list memory experiments is much less strong than it appears, and that, even if these effects do exist, there is currently no adequate explanation of why they would occur.

5.1 List memory, concreteness, and Dual Coding Theory

In a list memory task, participants read or hear a list of words one after the other, and are then required to reproduce as many of those words as possible after the list has finished being delivered. A list memory task can either require participants to preserve the order of items in the list, in which case it is known as a 'serial recall' task, or it can allow participants to reproduce items in any order, in which case it is known as a 'free recall' task. The rationale behind subjecting participants to these tasks is that if lists with certain properties are easier to remember than lists with different properties, then this differential performance speaks to some fact about the nature of memory: the reason that list A was easier to remember than list B might have important consequences for theories of cognition, based on their different properties. In some list memory concreteness experiments, participants are given instructions that are designed to bias them towards certain mnemonic strategies

(Paivio et al., 1994). For example, participants might be told explicitly to try and generate a mental image of the referent of each word in a list. In this chapter, we are interested in *default* concreteness effects in list memory, where participants are simply told to try and remember a list of words without any other guidance. As we saw in Chapter 2, historically the most popular theory of conceptual processing based on the concreteness-abstractness construct is Dual Coding Theory (Paivio, 1986). DCT posits a qualitative and quantitative representational difference between concrete and abstract concepts based on the assumption that there are (at least) two types of mental representation (imagens and logogens), which feature in two types of cognitive process (the non-verbal and verbal codes). DCT holds that differential performance on concrete versus abstract stimuli should only consistently occur when: either a) participants are instructed to approach the experiment in such a way that this representational difference is likely to manifest, or b) the task is constructed in such a way that it is impossible for this representational difference *not* to manifest no matter how a participant approaches it. My use of the word 'default' is, therefore, meant to capture the idea that, in the studies discussed below, concreteness effects supposedly emerge even in list memory paradigms that do not overtly encourage a processing strategy that would explain why concrete items show an advantage over abstract items.

Default concreteness effects in list memory tasks, whereby lists of concrete items are easier to recall than lists of abstract items, are somewhat consistent, and when such an effect is obtained, the explanation of the effect is slightly different to that of DCT. For example, Romani et al. (2008, p. 313) suggest that 'lexical semantic' information facilitates task performance: words with high concreteness values are easier to remember than words with low concreteness values because high concreteness value words tend to have 'richer semantic representations'. The claim here is that concrete concepts are constituted by representations that are 'richer' than those constitute abstract concepts, and that this property of richness facilitates recall. We shall return to this claim in the discussion section because its seeming innocuousness belies a troubling explanatory gap, and it is not quite the same as the claims made by DCT. In DCT, it is not the inherent 'richness' of concrete items that gives them their advantage over abstract items on behavioural measures. Instead this advantage is attributed to a type of *processing strategy* that is likely to benefit from the structural properties that concrete items tend to have relative to abstract items. The important point is that this advantage does not occur *by default* from the point of view of DCT: under DCT, concrete items do not have properties that make them

intrinsically easier to process than abstract items just by virtue of having those properties. One previously popular position, Context Availability Theory (CAT), does hold that this advantage should occur by default. However, the architects of this theory have themselves reported that DCT offers a better account of experimental data (Schwanenflugel et al., 1992). We shall return to a more in-depth discussion of DCT later on, where it becomes relevant to the issues raised by the experimental aspects of this chapter.

Let us now briefly consider the prevalence of default concreteness effects in the list memory literature. One very early study reports no difference between concrete and abstract items (Brener, 1940). Romani et al. (2008) cite Paivio and Csapo (1969) as finding no concreteness effects in list memory tasks despite the fact that one of Paivio and Csapo's experiments can be interpreted as providing evidence *for* concreteness effects. Romani et al. (2008) themselves report a large number of list memory experiments, both free and serial recall, that generally show strong concreteness effects. Miller and Roodenrys (2009) report a series of experiments designed to investigate potential interactions between the effects of concreteness and word frequency in serial recall. Their results indicate concreteness effects at both high and low word frequencies, and an interaction between frequency and concreteness such that concreteness effects are enhanced when words are of low frequency. Walker and Hulme (1999) and Allen and Hulme (2006) also each report a series of concreteness effects in serial recall. Finally, Morr (1981) tested participants' recall of high and low imagery words. Imagery and concreteness are separate but correlated variables. Morr (1981) found no advantage for high imagery (likely to be highly concrete) versus low imagery (likely to be abstract) items. Despite some inconsistencies then, it would seem that default concreteness effects do frequently emerge in list memory tasks.

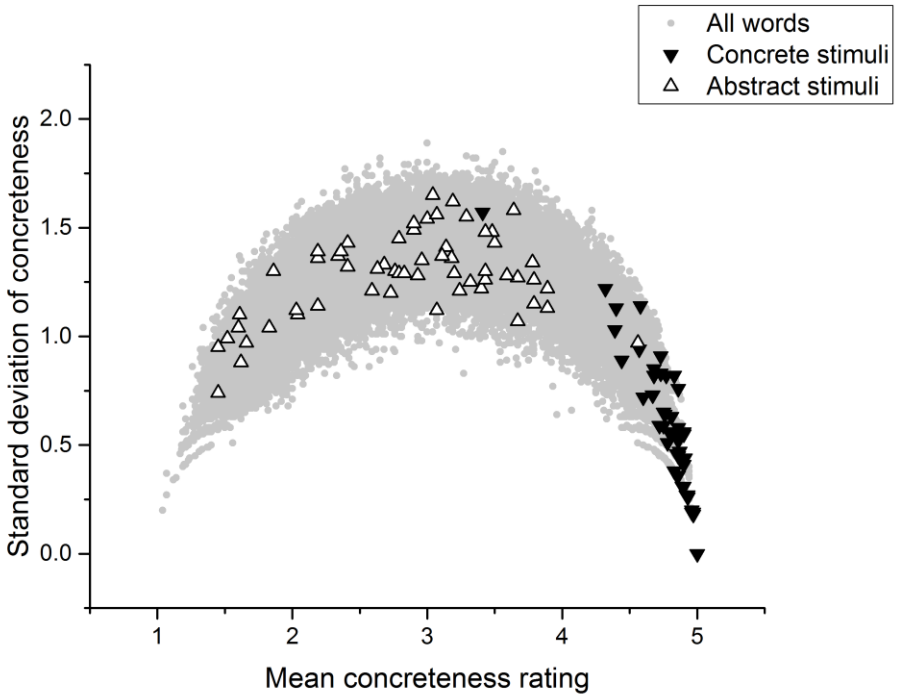
5.2 Methodological issues with list memory concreteness experiments

There are some methodological problems present in varying degrees that could *potentially* explain the increased recall of concrete items over abstract items in the studies just mentioned. It should be stressed here that the intention is not to argue that concreteness effects in list memory tasks are definitely attributable to other factors, or that concreteness effects in list memory tasks do not occur. Rather, the aim is to highlight the fact that the interpretation of the results of list memory studies of concreteness to date is made difficult by these factors, and it would presumably be

a good thing if we could remove them. The main issue is that the midscale stimulus problem identified in the previous chapter applies to all of the experiments we will consider here. There are also various experiment-specific stimuli properties that potentially bias recall rates in favour of one experimental condition over another, and that also have nothing to do with concreteness. We shall now consider these problems in turn.

Figures 5-1 to 5-4 below depict the stimuli featured in list memory experiments. All figures plot the mean concreteness ratings and standard deviations of concreteness ratings of every word in the Brysbaert et al. (2013) database. Each individual figure plots the mean ratings and standard deviations of the abstract and concrete stimuli featured in Romani et al. (2008), Allen and Hulme (2006), Miller and Roodenrys (2009), and Walker and Hulme (1999), respectively, over this whole distribution. The stimuli featured in Paivio and Csapo (1969) are available but not graphed because only 9 words in each condition featured in their experiments. This small stimuli pool generates its own problems and this will be discussed below. However, the criticisms I am about to level at the stimuli featured in Romani et al. (2008), Walker and Hulme (1999), Miller and Roodenrys (2009), and Allen and Hulme (2006) also apply to the stimuli featured in Paivio and Csapo (1969). The stimuli featured in Morr (1981) are not available for analysis. Note that although the experiments that I report below feature nominal stimuli, and most studies under discussion here also featured nouns, occasionally their stimulus sets did feature other word classes alongside nouns. In the case of Allen and Hulme (2006), many abstract items were not nominal. Therefore, in order to display the maximum number of stimuli for all experiments, I plot the entire Brysbaert et al. (2013) database instead of just the nominal subsection of it.

Figure 5-1 Stimuli featured in Romani et al. (2008)



We saw in the previous chapter what the issue with these stimuli is. The problem is that is not clear that the stimuli that make up the abstract condition in Romani et al.'s experiments are actually abstract. The concrete words tend to have low standard deviations, whereas the abstract stimuli tend to have high standard deviations and are drawn from the middle of the scale, rather than the unequivocally abstract half of the scale. This is potentially problematic for the validity of Romani et al.'s (2008) conclusions regarding concreteness effects because the stimuli that made up their abstract stimuli were not actually abstract.

Figure 5-2 Stimuli featured in Allen and Hulme (2006)

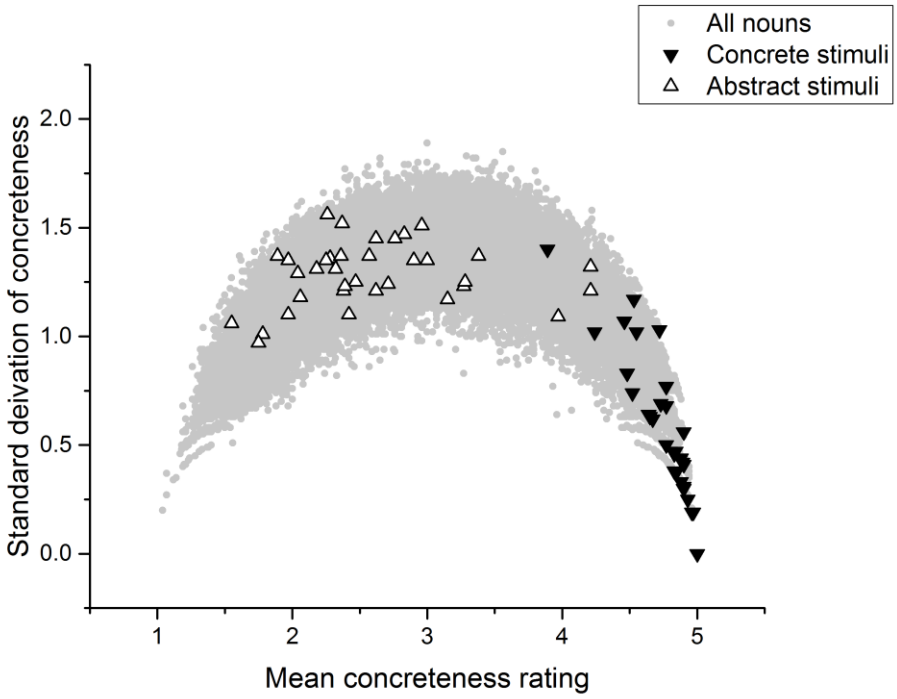
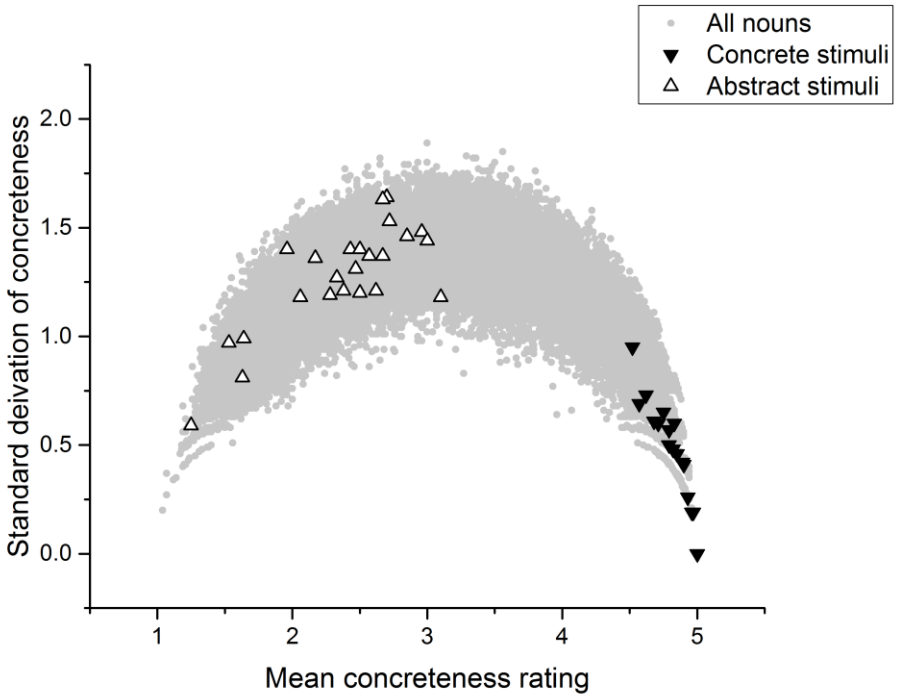


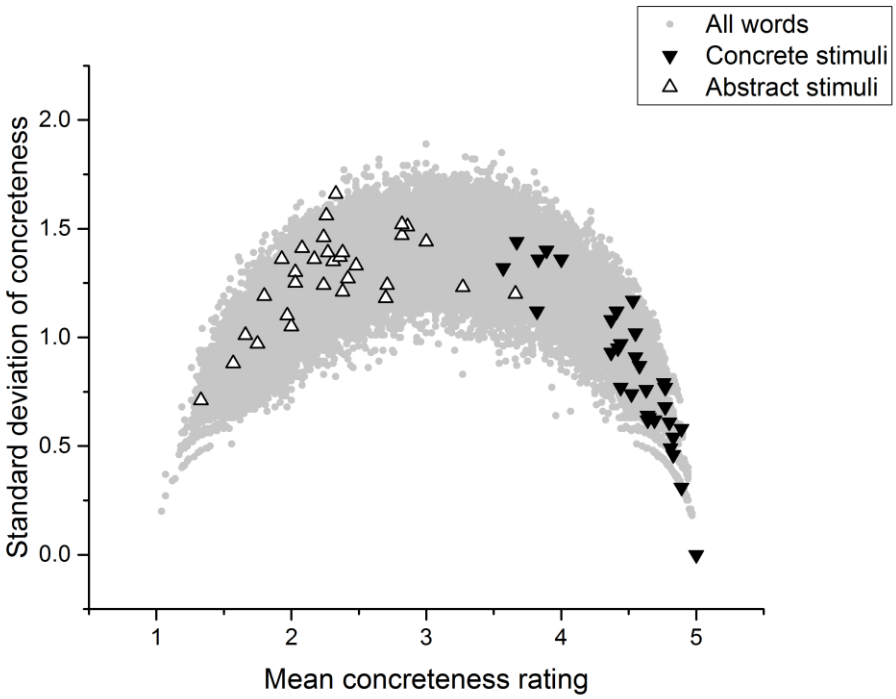
Figure 5-2 depicts the abstract and concrete stimuli featured in Allen and Hulme (2006). We can see that the abstract-midscale disagreement criticism does not apply quite so strongly to these items, but the issue is still present. That is, there are many 'abstract' stimuli here that have standard deviations well above 1, indicating that people disagreed about whether they were abstract in the first place. The range of mean ratings of concreteness for the abstract condition is also clearly much higher than for the concrete condition. Once again, a relatively homogenous group of concrete words has been compared to a heterogeneous group of words about which participants tended to disagree. Figure 5-3 below plots the stimuli featured in Miller and Roodenrys (2009):

Figure 5-3 Stimuli featured in Miller and Roodenrys (2009)



Again, there is a marked difference in standard deviations between the concrete stimuli and the abstract stimuli. Furthermore, the standard deviations of the abstract stimuli are so high (well above 1 in the majority of cases) that the mean value does not reflect the judgments that participants were actually making. Finally, consider Figure 5-4, which depicts the stimuli featured in Walker and Hulme (1999):

Figure 5-4 Stimuli featured in Walker and Hulme (1999)



The midscale criticism perhaps applies least to this set of stimuli, although it is still clearly the case that the concrete stimuli tended to have lower standard deviations than the abstract stimuli. The reasons that this is concerning have already been expounded. The upshot of this is that it is not clear that these studies actually provide evidence for concreteness effects. The reason is that the comparison being made was *meant* to be between concrete and abstract items, but the comparison that was *actually* made was between concrete items on the one hand, and a group of stimuli about which participants disagree on the other. It could be the case that words that engender disagreement are those words that are hard to remember, and that this explains processing differences that were previously attributed to concreteness/abstractness. My experiments reported below were designed to test this possibility.

I now consider issues that are specific to individual studies. Paivio and Csapo's (1969) first set of experiments featured a very small number of stimuli: 9 words per condition. Paivio and Csapo (1969) do not model item variability and so this issue is especially relevant. In Allen and Hulme (2006), the majority of the concrete stimuli were unambiguously nominal by word form, whereas the majority of abstract stimuli were either ambiguous between nominal and verbal parts of speech,

or unambiguously verbal. This difference was compounded by the fact that *before* the list memory task was conducted, every participant took part in a speaking-to-definition task that required them to conceive of most of the concrete stimuli as nouns, and most of the abstract stimuli as verbs. That is, the definitions provided for the concrete stimuli biased a nominal interpretation, whereas the definitions provided for the abstract stimuli tended to bias a verbal interpretation (generally by beginning the definition with the infinitive particle 'to'). This is an issue because the extent to which noun processing differs from verb processing has been the subject of fierce debate for decades. Vigliocco et al. (2011) provide a recent review in which they argue that once 'semantic' factors have been taken into account, much of the brain imaging evidence for a *neuroanatomical* difference in noun-verb processing disappears. However, they still note that 'verb processing imposes greater demands than noun processing in most cases' (Vigliocco et al., 2011, p. 422). Given that participants were biased to consider many of the abstract stimuli as verbs, and many of the concrete stimuli as nouns, it *could* be that the increased recall rate for the concrete condition in Allen and Hulme (2006) is really the result of the fact that verbs are generally harder to process than nouns. Once again, the point here is not that this confound definitely explains the concreteness effects that Allen and Hulme (2006) find. The point is rather that when confronted with a combination of numerous issues that might affect performance across conditions, we can be less sure that performance advantages in the concrete condition should be attributed solely to the concreteness of the stimuli presented. Finally, there is a confound present in Romani et al. (2008), Walker and Hulme (1999) and Miller and Roodenrys (2009) that might actually have supressed an advantage for concrete items over abstract items. In all three studies, the abstract stimuli had a significantly higher Age of Acquisition (AoA) than the concrete stimuli. Confirmatory independent-samples t-tests are summarised in Table 5-1:

Table 5-1 - Age of acquisition of stimuli in previous experiments

Experiment	Concrete mean	Abstract mean	T-statistic, df, p-value	Lower 95% CI	Upper 95% CI
Romani et al.	5.84	8.44	-8.52, 132, p<0.0001	-3.2	-1.99
Walker and Hulme	6.67	8.21	-3.427, p=0.001	-2.44	-0.64
Miller and Roodenrys	4.62	7.03	p<0.0001	-3.22	-1.59

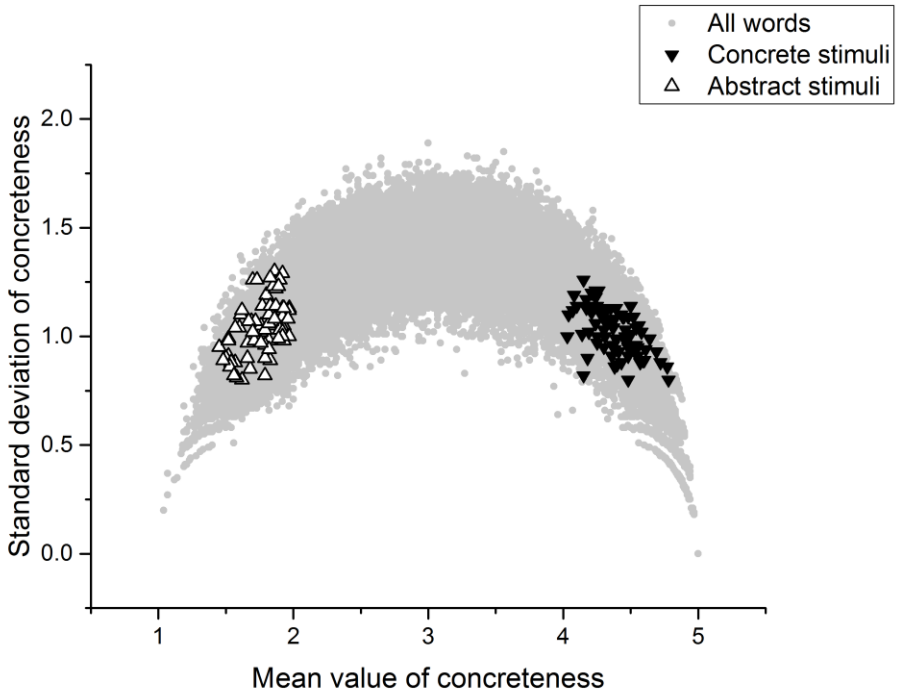
Kuperman et al. (2012) provide an Age of Acquisition norm database for 60,000 English words and summarise evidence showing that AoA should be considered a major factor in single word processing paradigms alongside characteristics such as word frequency and word length. Morr (1981) presents data suggesting that the *higher* the AoA of a word, the easier it is to recall in list memory tasks. It should be noted that the stimuli featured in every experiment discussed in the current section had relatively high frequencies and relatively low AoAs compared to the stimuli featured in the experiment I report below. Morr (1981, p. 281) explicitly confines his explanation of his reported AoA effect (that high AoA words 'facilitate distinctive encoding' relative to low AoA words) to words with relatively high frequencies. In any case, controlling for AoA in the present experiment should eliminate this nuisance effect from consideration.

We have seen that there are various issues with the list memory studies under discussion. These issues reduce certainty in the finding that concrete words are, by default, easier to remember than abstract words. Most importantly, it is not clear that concrete words were compared to truly abstract words in the first place. There are also some potentially relevant confounds in various stimuli sets employed that could account for at least some of the differential performance in favour of concrete word lists over abstract word lists. I will now report three new experiments that, hopefully, avoid the issues discussed throughout this section. These new experiments, therefore, provide a clearer test of default concreteness effects in list memory tasks.

5.3 Experiment 1

The purpose of this experiment was to replicate an experiment reported in Romani et al. (2008) while controlling for the potentially problematic confound between the mean value of concreteness rating and the standard deviation of that rating, as well as the other stimuli confounds present in the other list memory studies discussed above. Romani et al. (2008) presented participants with lists of words, and asked them to recall words from a list immediately after the presentation of the last word of that list. They report a range of experiments designed to investigate the effects of different task demands and types of representation on task performance. Romani et al. report that participants were significantly better at recalling lists of words that consisted entirely of concrete words versus lists that consisted entirely of abstract words. The focus of the new experiment reported here is on investigating the reliability of this concreteness effect when the standard deviations of the concreteness value of words across lists is controlled, while also directly manipulating this standard deviation in order to ascertain whether the standard deviation itself has a significant effect on task performance. Figure 5-5 below plots the mean concreteness values and standard deviations of concreteness of the concrete and abstract stimuli used in the present experiment in the same way that the stimuli used in previous experiments were plotted in the previous section.

Figure 5-5 Stimuli featured in experiment 1



We can see that the contrast in concreteness between conditions is maximised, and that the difference in the standard deviations of concreteness rating is controlled. Of interest is whether the concreteness effect still occurs when these new controls are enforced. Romani et al. (2008) report a range of list memory tasks that employ various manipulations. The specific experiment replicated here is Experiment 3B, which is a free recall task in which participants simply try and recall any word from the list that they can, regardless of order. This experiment was chosen over the serial recall tasks (in which participants must recall the words in the order they were presented) because Romani et al. (2008) report that concreteness effects are stronger in free recall tasks than in serial recall tasks. A free recall task, therefore, provides the most robust test of the concreteness effect. An additional two experimental conditions were added: agreement and disagreement conditions. Words in the agreement condition were taken from the middle of the scale and had relatively low standard deviations. Words in the disagreement condition were taken from the middle of the scale and had relatively high standard deviations. A comparison of these two new conditions with concrete and abstract conditions will help answer the question of whether midscale variability in concreteness ratings indexes a psycholinguistically relevant effect. In this way, the importance of the midscale problem outlined in the previous chapter can be assessed.

Participants

Originally, 60 native speakers of English with no reported neurological disorders were recruited from the University College London SONA psychology pool. Of these, 50 managed to complete the experiment (the other 10 either did not turn up or cancelled their session). 35 of the participants were male; mean age = 23.7 (7.3). All participants were either awarded course credit or paid £6 for their time.

Materials

Forty lists each containing 8 words were generated. There were 4 experimental conditions and each experimental condition comprised 10 lists. Stimuli were controlled for the following psycholinguistic variables: standard deviation of concreteness rating, frequency, age of acquisition, number of phonemes, number of letters, number of syllables. Table 5-2 below contains the mean values (standard deviations in parentheses) of each of these variables for each condition:

Table 5-2 Properties of stimuli featured in experiment 1

Condition	Mean concreteness	SD concreteness	AoA	Zipf frequency	L Phon	Length	Syll
concrete	4.38 (0.17)	1.02 (0.11)	10.45 (2.05)	3.34 (0.79)	5.59 (0.94)	6.93 (1.06)	2.00
abstract	1.78 (0.14)	1.04 (0.12)	10.58 (2.09)	3.38 (0.83)	5.56 (0.93)	6.84 (1.17)	2.00
agree	3.17 (0.7)	1.08 (0.07)	10.09 (1.9)	3.15 (0.85)	5.63 (1.03)	6.93 (1.21)	2.00
disagree	3.1 (0.36)	1.65 (0.05)	10.23 (2.04)	3.13 (0.81)	5.76 (1.10)	6.9 (1.32)	2.00

Table legend:

Mean concreteness: Mean concreteness rating
SD concreteness: The mean standard deviation of the concreteness ratings
AoA: Age of acquisition
Zipf frequency: Word frequency in Zipf units
L Phon: Length of word in phonemes
Length: Length of word in letters
Syll: Number of syllables

Psycholinguistic variable information was gathered from Brysbaert et al. (2013), Kuperman et al. (2012b) and the English Lexicon Project (Balota et al., 2007). Word data was combined into a single master database. This master database was then searched using the MATCH program in order to create stimuli lists that near-optimally maximised the contrasts of interest while minimising the difference in nuisance variables across conditions (Van Casteren and Davis, 2007). The four conditions were: concrete, abstract, agreement, disagreement. Concrete lists contained words that had mean values between 4 and 5 on the Brysbaert et al. (2013) concreteness scale. Abstract lists contained words that had mean values between 1 and 2 on the Brysbaert (2013) concreteness scale. Agreement and disagreement lists contained words that had mean values between 2.5 and 3.5 on the Brysbaert et al. (2013) concreteness scale. Concrete, abstract, and agreement lists were constructed such that the standard deviations of the concreteness ratings of the words in those lists were similar. Regarding the concrete-abstract condition comparison, there was no difference in the standard deviation of concreteness rating (1.02 vs 1.04), but there was a difference in the mean rating of concreteness (4.38 vs 1.78). Therefore, the contrast in mean rating was maximised while contrast in standard deviation was minimised. Regarding the agree-disagree comparison, there was no difference in the mean rating of concreteness (3.17 vs 3.1), but there was a difference between the two conditions in the mean standard deviation of concreteness rating (1.08 vs 1.65). Therefore, the mean value of concreteness rating was held constant in this comparison while a contrast in standard deviation of concreteness rating was introduced. Table 5-3 below contains a sample list from each condition, and full lists of stimuli featured in all experiments reported in this study are included in Appendix A:

Table 5-3 Example stimuli from experiment 1

Condition	
Concrete	Beaker, Clinic, Tango, Clothing, Amber, Jackal, Roulette, Survey
Abstract	Desire, Mystique, Intent, Vantage, Glory, Nuance, Unease, Motive
Agree	Diesel, Roughhouse, Attempt, Whiner, Viewpoint, Freshness, Stampede, Leader
Disagree	Slipstream, Audit, Poorhouse, Minute, Rival, Tribune, Abyss, Spectrum

Procedure

The experimenter read all of the words from a list one after the other. There was a two second pause in between each word being read out. The order of the lists and the order of the words within each list were randomised for each participant. After the experimenter finished reading out a list, the participant spoke out loud any and all words that they could remember from that list. The experimenter recorded every word that the participant spoke. Because this was a free recall task, the order in which participants recalled the words did not matter. Participants were not penalised for making errors or substitutions, or for saying a word that had not actually been in the list. The experiment lasted approximately 35 minutes. The maximum score for a list was 8, because there were 8 words in each list. Each individual list was counted as an item in the statistical analyses.

Results

Table 5-4 below summarises the mean number of words remembered (and standard deviations) by condition.

Table 5-4 Mean words recalled by condition for Experiment 1

Condition	Mean words recalled (SD)	Mean percentage recalled
Concrete	4.67 (1.35)	58.4%
Abstract	4.48 (1.24)	56%
Disagree	4.38 (1.28)	54.6%
Agree	4.45 (1.35)	55.6%

It is immediately apparent that the mean number of words remembered in the current experiment is lower than the mean number of recalled words reported in Romani et al. (2008) (83.3% for concrete lists; 70.3% for abstract lists compared to 58.4% and 56% respectively). This may seem troubling because the aim here is replicate Romani et al.'s experiment, and differing levels of performance might suggest some error in the application of the methodology. However, there is a ready explanation for this discrepancy: Romani et al. ran two experiments, Experiment 3A and Experiment 3B, one after the other on the same participants. Experiment 3A was an order reconstruction task in which participants were presented with words on pieces of card which were then shuffled. The stimuli in Experiment 3A (and, presumably, the word cards) were the same as those used in Experiment 3B. Participants had therefore just encountered the stimuli they were required to remember in Experiment 3B in a previous experiment.

The results were analysed with a mixed effects model in R using the lme4 package (Bates et al., 2015). The lmertest package was used in order to obtain p-values for the comparisons of interest via Satterthwaite Approximation (Kuznetsova et al., 2015). The mixed effects model examined the fixed effect of experimental condition on the number of words remembered per trial, with subjects and items being treated as random effects with varying intercepts.

The statistical contrasts were abstract, disagreement, and agreement conditions versus the concrete condition; that is, a treatment contrast, with the concrete condition representing the baseline condition. Table 5-5 below displays the results of this analysis:

Table 5-5 Summary of mixed effects model for Experiment 1

Fixed effects	Effect estimate	Error	df	t	p	Lower 95%CI	Higher 95%CI
Abstract	-0.19	-0.12	39.25	-1.56	0.13	-0.43	0.05
Agree	-0.22	-0.12	39.25	-1.79	0.08	-0.46	0.03
Disagree	-0.29	-0.12	39.25	-2.42	0.02	-0.54	-0.05

Because three non-independent hypothesis tests were run on the same data, a Bonferroni correction was applied. Assuming a conventional alpha level of 0.05, the corrected alpha level is therefore 0.05/3 = 0.017. The concrete-abstract contrast was not statistically significant ($p = 0.13$). Therefore, there was no evidence for an advantage for concrete word lists over abstract word lists, contrary to the findings in Romani et al. (2008), Walker and Hulme (1999), Allen and Hulme (2006), and Miller and Roodenrys (2009). We might immediately wonder whether the results of this replication experiment are aberrant in some way: the previous four studies have all shown consistent concreteness effects over multiple experiments, and this new replication is the only experiment where this effect is absent. I have two responses to this suggestion. Firstly, in terms of effect sizes, the results across all studies are actually quite similar. No measure of effect size or 95% confidence intervals are presented in any of the previously conducted list memory experiments under discussion, so a formal comparison of effect size is not possible, but the results are not wildly discrepant: differences of less than a word on average across conditions are common. Table 5-6 below summarises the mean differences in recall across concrete and abstract conditions in various list memory experiments.

Table 5-6 Mean differences in words remembered in concrete and abstract conditions in various list memory experiments

Study	Experiment	Mean concrete words recalled	Mean abstract words recalled	Difference in means
Present experiment	n/a	4.76	4.48	0.28
Romani et al.	3B	6.8	5.6	1.2
Allen and Hulme	1	4.2	3.7	0.5
Miller and Roodenrys	1	3.6	3.1	0.5

Note that the list length varied across these four experiments from 6-word lists to 8-word lists, and the second two experiments featured in the table were serial recall and not free recall. In the reporting of list memory experiments, the difference in concrete and abstract conditions is generally expressed in terms of a percentage. But these percentage formats might confuse the interpretation of these results somewhat. If the mean percentage of concrete words recalled from an 8-word list was 83.3%, as is the case in Romani et al. (2008), that means 6.8 concrete words were recalled on average. If the mean percentage of abstract words recalled from an 8-word list was 70.3%, that means 5.6 abstract words were recalled on average. When these differences are expressed in terms of means rather than percentages, as is the case in table 6, the differences between conditions seem less dramatic. A 13% difference is really a difference of 1.2 words on average in an 8-word list experiment. Also, the results of the present experiment seem less divergent. In Allen and Hulme (2006) and Miller and Roodenrys (2009), the mean difference in recall across conditions was just 0.5 words. In the present experiment, it was 0.28 words, meaning an informal difference in effect of just 0.22 words on average.

I now move to a discussion of the other two experimental contrasts. Neither contrast was statistically significant at the Bonferroni-corrected alpha-level (concrete versus agreement $p = 0.08$, concrete versus disagreement $p = 0.02$). There was therefore no evidence that it is simply words from the middle of the concreteness scale that are harder to remember than words from the extreme concrete end of the scale, and there was no evidence that words with high standard deviations in rating are harder to remember than words from the extreme concrete end of the scale.

However, it is important to note that experiment 1 suffers from a lack of power because there are only 10 items per condition. This could be the reason that no statistically significant results were obtained. In order to account for this possibility, the data were reanalysed using a Bayesian model comparison analysis in the BayesFactor package for R (Morey et al., 2015) with the default settings and priors. If the results of the frequentist analysis presented in the preceding paragraphs were due to low power, then the Bayes Factors produced by this analysis are likely to be between 1/3 and 3, which would indicate that the data do not decide the issue either way.

Kruschke (2011, p. 310) argues that the Bayes Factor generated from a model comparison analysis of an experimental design with multiple conditions may be misleading for various reasons. Therefore, the total results dataset of experiment 1 was partitioned into three smaller datasets that reflected the pairwise comparisons of interest between the conditions: one concrete/abstract comparison, one concrete/agree comparison, and one concrete/disagree comparison. In every case, a model including a parameter for the fixed effect of condition was compared to a null model that featured only subjects and items as random effects. The resulting Bayes Factors for each comparison were:

Concrete vs abstract: 0.32

Concrete vs agree: 0.38

Concrete vs disagree: 0.66

For the concrete-abstract comparison, there is positive evidence in favour of a null effect (BF = 0.32), according to the conventions discussed in (Wagenmakers, 2007). For the other two comparisons, the Bayes Factor indicates that the data do not decide between the null or alternative models. Taken together with the frequentist analysis presented previously (all p-values above the threshold for statistical significance), these results suggest that there really was no difference in recall between concrete and abstract conditions. However, the evidence for a null difference in the other comparisons is inconclusive.

Before moving on to the second replication experiment, I shall note some shortcomings of experiment 1 that were remedied in experiment 2. Firstly, the standard deviations of the concreteness ratings of both the concrete and abstract stimuli were relatively high: above 1 in many cases. It could be that, given the

concerns raised in previous sections, neither condition provided an accurate sample from the truly concrete or abstract sections of the scale. In the second experiment that I am about to report, the standard deviations of the conditions were more tightly constrained so that in the concrete and abstract conditions, all standard deviations were below 1. Secondly, the design of experiment 1 required that the researcher be in constant contact with the participant. Throughout the experiment, the researcher read out words and recorded the participant's responses. In an ingenious series of experiments, Intons-Peterson (1983) shows that so-called 'demand characteristics', whereby a researcher can unwittingly affect a participant's behaviour, can exert a huge influence over the results of imagery/concreteness experiments. Given that the expectation before the current replication experiments were run was that the concreteness effect should *increase* rather than disappear, this should not be too great a concern. Nevertheless, the second experiment was run online with no researcher-participant contact whatsoever in order to eliminate the possibility that this factor may have influenced the results of the experiment. I now turn to a report of this second replication attempt, which was a paired associate learning task.

5.4 Experiment 2

In paired associate learning, participants are presented with pairs of words, one after the other, and their task is to remember as many of the words as possible (Paivio et al., 1994). Paivio et al. (1994) presented participants with lists consisting of both concrete and abstract word pairs, and report that concrete word pairs were recalled better than abstract word pairs. This effect has been obtained in many list learning and paired associate learning experiments (Begg, 1972; Nelson and Schreiber, 1992; Paivio et al., 2000, 1994). Paired associate learning is therefore a good candidate for examining the midscale problem outlined in previous sections.

Paivio et al. (1994) employed a range of different manipulations across two experiments, but in this replication I focus on the simplest version of this paradigm, which is a free recall task. In free recall, after the list of word pairs has been presented, a participant simply has to write down any and all words that they remember from the entire pool of words that they have seen. Note that although there is some disagreement about how strong concreteness effects are in paired associate free recall (Marschark and Hunt, 1989), generally studies do find the effect. The results of a free recall task are also more comparable to the results of experiment 1 above, because it was also a free recall task. The aim of the present experiment was to test whether a concreteness effect still occurs if the contrast between concrete and

abstract stimuli is maximised and the standard deviations of their concreteness scores are controlled. In addition to the concrete and abstract conditions featured in the paired associate learning studies mentioned in this section, the present experiment also included a midscale condition in order to provide a second test of the hypothesis that high-standard deviation midscale words are harder to remember than words from the concrete end of the concreteness scale.

Participants

Sixty native speakers of English with no reported neurological disorders were recruited from the Prolific Academic website. 38 participants were male; mean age = 31.7 (10.0). All participants were paid £6 for their time.

Materials

Figure 5-6 depicts the means and standard deviations of the concreteness ratings for the concrete and abstract stimuli in experiment 2:

Figure 5-6 Concrete and abstract stimuli featured in experiment 2

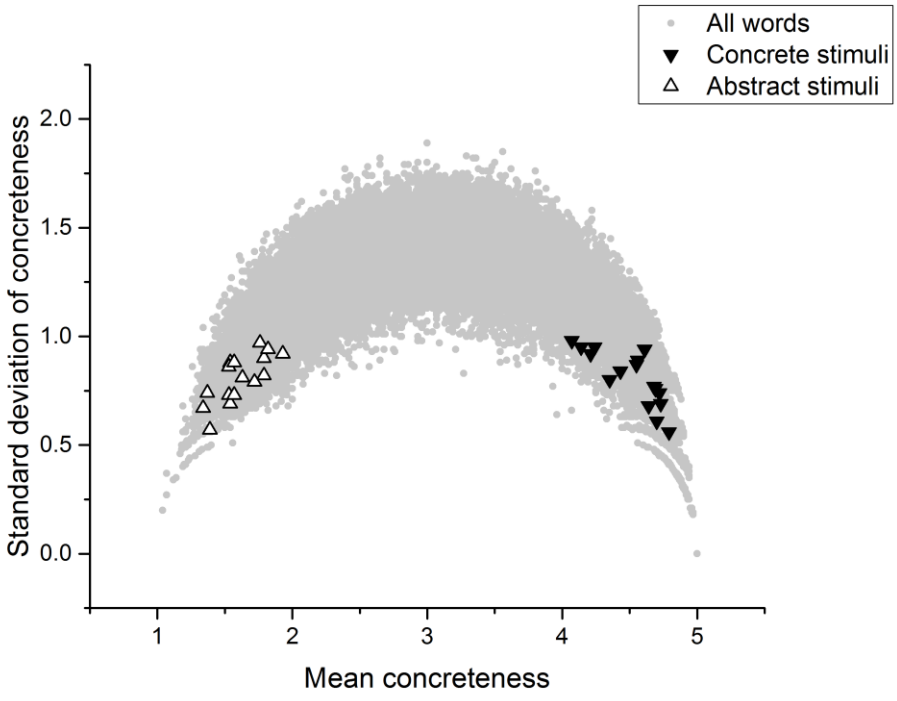


Table 5-7 below displays the psycholinguistic characteristics of the stimuli featured in experiment 2 by condition.

Table 5-7 Properties of stimuli featured in experiment 2

Condition	Mean concreteness	SD concreteness	AOA	Zipf freq	Phon	Syll	Length	BG mean
concrete	4.51 (0.23)	0.91 (0.13)	9.92 (1.9)	3.54 (0.56)	4.75 (0.2)	1.75 (0.43)	6.125 (1.41)	3573 (1151)
abstract	1.61 (0.17)	0.81 (0.11)	10.04 (1.64)	3.48 (0.69)	5.25 (1.44)	1.75 (0.43)	6.44 (1.5)	3457 (1176)
disagree	3 (0.23)	1.33 (0.02)	9.78 (1.95)	3.72 (0.78)	5.75 (1.48)	1.81 (0.39)	6.38 (1.45)	3218 (957)

Table legend
Mean concreteness: Mean concreteness rating
SD concreteness: The mean standard deviation of the concreteness ratings
AoA: Age of acquisition
Zipf freq: Word frequency in Zipf units
Phon: Length of word in phonemes
Syll: Number of syllables
Length: Length of word in letters
BG mean: Mean bigram frequency

As in experiment 1, concrete words were only considered concrete if they were 4 or above on the Brysbaert et al. (2013) concreteness scale, and abstract words were only considered abstract if they were 2 or below on this scale. In experiment 2, the additional control variable of mean bigram frequency was introduced, because participants would be reading and writing words as opposed to hearing and speaking them. Bigram frequency is a measure of orthographic regularity that has been regularly shown to have an impact on single word processing (Kuperman et al., 2012b). There were eight pairs of words in each condition, and therefore there were 16 words total in each condition and 24 critical item pairs overall.

Procedure

Participants undertook the experiment online via a Qualtrics survey distributed over the Prolific Academic service. Participants were presented with pairs of words, one

after the other. Following Marschark and Hunt (1989) and Paivio et al. (1994), each pair of words was presented on the participant's computer screen for 8 seconds. There were 8 pairs in each of the three conditions, and all pairs were presented in a randomised non-blocked order for each participant. The ordering of the words in each pair from left to right on the computer screen was not randomised. At the beginning and end of the list, 3 pairs of filler items were included in order to soak up primacy and recency effects. Participants also received a short practice trial with words not included in the main experiment in order to ensure that they understood the task, and that their computers and internet connections were working properly. Once the list of pairs was finished, participants could type out any and all words that they remembered from the list. Once they were finished, they pressed a 'submit' button that ended the experiment. There were three experimental conditions: a word pair could either consist of concrete, abstract, or midscale 'disagreement' items. The experiment lasted approximately 15 minutes. The likelihood of remembering a single word was the dependent variable, and so individual words constituted the items in this experiment.

Results

Table 5-8 below displays the mean number of words remembered across conditions in experiment 2.

Table 5-8 Mean words recalled by condition in experiment 2

Condition	Mean words recalled	Mean percentage recalled
Concrete	3 (2.73)	18.6%
Abstract	3.43 (3.07)	21.5%
Disagree	3.05 (2.84)	19.1%

We can immediately note a number of things. Firstly, the number of words recalled out of 16 was low but the variability across participants was large, as indicated by the high standard deviations of the mean number of words recalled: with standard deviations and means of approximately 3, many participants must only have responded with 1 or 2 words per trial. This could have been caused by two factors. It could be that the task was difficult, and so we are seeing floor effects for some participants. Alternatively, some participants may just not have been attending to the task properly. This makes the results of this experiment difficult to interpret because

there is no way of deciding between these two possibilities, and therefore no way of knowing which participants to exclude from statistical analyses. In any case, the mean number of words in the abstract condition was numerically larger than that of the concrete condition (3 < 3.43), and so already we have failed to find evidence in favour of concrete stimuli advantage in paired associate learning. Finally, the differences between the means of concrete and disagree conditions are miniscule (3 versus 3.05 respectively).

The data were analysed using a generalized linear mixed model fit by maximum likelihood (Laplace Approximation) using the glmer function from the lme4 package in R. The dependent variable in this analysis was therefore the likelihood of a participant recalling a word[8]. Individual words were treated as items. Subjects and items were included as random effects with varying intercepts, and the fixed effect of condition was the effect of interest. Both abstract and disagree conditions were compared to the concrete condition. The results of this analysis are presented in table 5-9:

Table 5-9 Summary of generalized linear mixed model analysis of experiment 2

Effect	Effect estimate	Std. Error	z	p
Abstract	0.19	0.15	1.3	0.2
Disagree	0.02	0.15	0.15	0.88

Experiment 2 generated no statistically significant effects: p = 0.2 for the concrete-abstract contrast, and p = 0.88 for the concrete-disagree contrast. This pattern of results is the same as that found in experiment 1: under conditions that should have made a concreteness effect stronger, a concreteness effect was not obtained. However, ultimately we should be cautious in drawing any conclusions from the results of experiment 2, because floor effects and/or participant disengagement may be obscuring any differences between conditions.

8 Analysing the data this way means that this experiment is arguably no longer a paired-associate learning task, presumably because it does not account for the paired relationship between words. In their free recall analyses, Paivio et al. (1994) calculate the proportions of words remembered and conduct by-subjects and by-items ANOVAs on these proportions. These analyses also ignore word-pair relationships and produce concreteness effects, and so I think we would still expect the analysis presented here to produce a concreteness effect.

Experiments 1 and 2 did not produce a concreteness effect. This is worrying given the concerns about the typically high standard deviations of abstract stimuli outlined above. If we increase a difference between conditions on some linear measure, then we would not expect experimental effects based on this measure to disappear. However, Kousta et al. (2009) show that words with a high emotional valence (whether positive or negative) enjoy a processing advantage over words with neutral emotional valance. Abstract words tend to be rated higher for emotional valance than concrete words, and this variable was not controlled in experiment 1 or 2. So it could be that there is a confound in the stimuli used in experiments 1 and 2 that obscured a concreteness effect. Warriner et al.'s (2013) emotional valance norms for ~14,000 English words allow us to check this possibility. Emotional valance is rated on a scale of 1 (highly negative) to 9 (highly positive), with a score of 5 indicating an emotionally neutral word. Given that either emotional positivity or negativity results in a processing advantage, the absolute value of 5 minus the emotional valance of a word provides a simple linear measure of emotional valance that ignores polarity (0 = totally neutral, 4 = highly emotionally valenced). Table 5-10 below presents the mean absolute emotional valences of the stimuli featured in experiments 1 and 2:

Table 5-10 Emotional valence of stimuli featured in experiments 1 and 2

Experiment	Concrete	Abstract	Disagree	Agree
1	0.82	1.17	0.88	1.15
2	0.91	1.61	0.99	N/A

In experiment 1, the emotional valence was more or less the same across concrete and abstract conditions. In experiment 2, the difference was larger, although as mentioned above experiment 2 may have been contaminated by floor effects in any case.

Another potential issue is that the words featured in experiments 1 and 2 were of relatively low frequency (between 3 and 4 on the Zipf scale), and so it could be that participants did not know all of the words. This could obscure any effect of manipulating concreteness. Brysbaert et al. (2013) provide a measure of how many of their participants reported that they knew a word. Table 5-11 below displays the mean percentage of participants who reported knowing a word for each condition in experiments 1 and 2.

Table 5-11 Mean percentage of participants who reported knowing words featured in experiments 1 and 2

Experiment	Concrete	Abstract	Disagree	Agree
1	98.5%	98.3%	97.7%	98.5%
2	99.5%	99.1%	98%	N/A

These percentages are high, and so it is likely that the number of participants in experiments 1 and 2 who did not know a word is very low. However, it would obviously be preferable if only words with known percentages of 100% were used. Unfortunately, for reasons detailed in the previous chapter, enforcing this control would mean that we would have less than 300 potential abstract stimuli to choose from. I now report an additional list memory experiment in order to provide a third test of concreteness effects in list memory.

5.5 Experiment 3

Experiment 3 was a free recall list memory experiment in the vein of experiment 1. There were three changes to the paradigm. Firstly, 6-word lists were used instead of 8-word lists. This change was made so that more trials per condition (15 in experiment 3 compared to 10 in experiment 1) could be fitted into roughly the same amount of time. Romani et al. (2008) and Miller and Roodenrys (2009) both report concreteness effects with 6-word lists. Secondly, the words were presented visually and participants wrote out the words at the end of a list instead of speaking them out loud. This change was made because, in order to maximise efficiency, the experiment was run over the internet using the Gorilla.sc platform. Finally, only three conditions were included: concrete, abstract, and midscale words with high standard deviations.

Participants

70 participants were recruited from the Prolific Academic website. Of these, 62 managed to complete the experiment. The other 8 did not respond to every trial, and so were excluded. 36 participants were male; mean age = 36.0 (11.7). The experiment was delivered via Gorilla.sc, and lasted approximately 35 minutes. Participants were paid £5 for their time.

Materials

Stimuli were controlled for the following psycholinguistic variables: standard deviation of concreteness rating, frequency, age of acquisition, number of syllables, number of

letters, mean bigram frequency, and emotional valence. Table 5-12 below contains the mean values (standard deviations in parentheses) of each of these variables for each condition, as well as the mean percentage of people in the Brysbaert et al. (2013) norms who reported knowing the words in each condition:

Table 5-12 Properties of stimuli featured in experiment 3

Condition	Mean concreteness	SD concreteness	AoA	Zipf freq	Syll	Length	BG mean	Absolute Valence	% known
Concrete	4.55 (0.17)	0.81 (0.12)	10.11 (1.28)	3.41 (0.48)	2.42 (0.86)	7.63 (1.79)	3649 (1134)	1.12 (0.77)	99%
Abstract	1.61 (0.15)	0.85 (0.11)	10.2 (1.95)	3.54 (0.72)	2.53 (0.89)	7.63 (1.95)	3710 (1208)	1.15 (0.78)	99%
Midscale	3.02 (0.26)	1.51 (0.77)	10.11 (1.99)	3.53 (0.72)	2.54 (0.86)	7.57 (1.89)	3737 (1184)	1.15 (0.77)	98.7%

Table legend:

Mean concreteness: Mean concreteness rating
SD concreteness: The mean standard deviation of the concreteness ratings
AoA: Age of acquisition
Zipf freq: Word frequency in Zipf units
Syll: Number of syllables
Length: Length of word in letters
BG mean: Mean bigram frequency
Absolute Valence: Absolute value of 5 minus the Warriner et al. (2013) emotional valence score
% Known: Percentage of participants who reported knowing a word in the Brysbaert et al. (2013) concreteness norms

There were three experimental conditions: concrete, abstract, and midscale. There were 15 six-word lists in each condition. As with experiment 1, each individual list was counted as an item in the statistical analyses.

Procedure

Participants were presented with words in sequence one at a time in the centre of their computer screens. As in Romani et al.'s (2008) visual paradigms, each word remained on the screen for 3 seconds. After each list had been presented, participants typed out any and all words that they could remember. They were told that the order of the words did not matter, and not to worry about spelling. Participants received two practice trials in order to ensure that they understood how to complete the experiment. The order of the lists and the order of the words within each list were randomised for each participant.

Results

Table 5-13 below summarises the mean number of words remembered (and standard deviations) by condition.

Table 5-13 Mean words recalled by condition for Experiment 3

Condition	**Mean words recalled (SD)**	**Mean percentage recalled**
Concrete	4.06 (1.31)	67.7%
Abstract	3.7 (1.25)	61.7%
Midscale	3.85 (1.28)	64.2%

The results from experiment 3 were analysed in the same way as the results from experiment 1. Both frequentist and Bayesian analyses are presented. Table 5-14 below displays the results of a mixed effects linear model with a fixed effect of condition and random intercepts for subjects and items.

Table 5-14 Summary of frequentist mixed effects model for experiment 3

Fixed effects	**Effect estimate**	**Error**	**df**	**t**	**p**	**Lower 95%CI**	**Higher 95%CI**

Abstract	-0.37	0.12	44.34	-3.11	0.003	-0.61	-0.13
Midscale	-0.21	0.12	44.34	-1.79	0.08	-0.45	0.03

After controlling for the effects of emotional valence, these results are more encouraging for the status of concreteness as a useful psycholinguistic variable. The concrete-abstract comparison is statistically significant at p = 0.003, and the difference is in the direction we would expect. The contrast between concrete and midscale conditions was not statistically significant (p = 0.08). Because this experiment still featured a relatively small number of items, a Bayesian model comparison analysis was deployed in an attempt to offset a potential lack of power. Again, the default settings and priors of the BayesFactor package were used. As with experiment 1, the results from experiment 3 were split into subsets so that the abstract and midscale conditions were compared to the concrete condition individually. The resulting Bayes Factors for each comparison were:

Concrete vs abstract: 5.85

Concrete vs midscale: 0.47

For the concrete/abstract comparison, the Bayesian analysis is comparable with the frequentist analysis. A model containing an effect of condition is 5.85 times more likely given the data than a model without this effect, which is positive in favour of a concreteness effect (Wagenmakers, 2007). Note however that the effect itself is small (a difference of 0.3 words on average). However, the concrete/midscale analysis was inconclusive. One thing to note is that experiment 3 featured words with similar rates of knowledge to those in experiments 1 and 2. Experiment 3 produced a concreteness effect, so this might partially allay any concern that experiments 1 and 2 produced null results because participants did not know the words used. I now turn to a general discussion of these results.

5.6 General discussion

The first two experiments did not produce a concreteness effect, but in Experiment 3 the typical concreteness effect re-emerged. At first glance, this might suggest that the potential problems with concreteness that I have outlined so far might not be too important. But, despite the fact that we found evidence for a concreteness effect in

one list memory experiment, I still think there are reasons to be cautious about interpreting this result.

Firstly, we need to consider what the explanation for this effect might be. The suggestion put forward by Romani et al. (2008) is that concrete representations are 'richer' than abstract representations in such a way that facilitates their recall. We might reasonably ask what it is for a concrete representation to be 'richer' than an abstract representation. DCT and related theories explain how it is that the representations that support concrete concepts differ from those that support abstract concepts, but they do so in such a way that they do not predict 'default' concreteness effects. Paivio has consistently held that concreteness effects should only emerge in specific paradigms in which the task demands bias participants to make use of one type of representation over the other, and that there are indeed certain tasks in which abstract language is likely to show an advantage over concrete language. In list memory tasks, if participants are explicitly instructed to construct a mental sensory simulation of the referent of each word as it is encountered, DCT predicts that their recall rates will be higher for concrete words than for abstract words because imagens lend themselves to mental sensory simulation whereas logogens do not. However, from the point of view of DCT, in neither Romani et al.'s (2008) experiments nor my replications was there any instruction that might bias participants to use processing strategies that relied differentially on imagens over logogens. It is always possible that some participants may have employed such strategies independently, but highly unlikely that they all would have done so. Indeed, Schwanenflugel et al. (1992) report a series of experiments that show how dependent concreteness effects are on task-specific factors and instructions (and note that they are admirably even-handed about these results given that they disconfirmed their own Context Availability hypothesis). So DCT does not necessarily predict a concreteness effect in the specific paradigm being discussed here.

Furthermore, it is not the case that DCT considers concrete concepts to be 'richer' than abstract concepts. It is therefore not clear that Romani et al. (2008) can rely on these frameworks in a definition of what constitutes representational richness. The only thing we seem to be left with is a folk appeal to intuitions concerning 'conceptual richness'. It is doubtful that this appeal can do the job as a plausible explanation for default concreteness effects. Consider the two nouns, 'doorknob' and 'spirituality'. According to the Brysbaert et al. (2013) norms, 'doorknob' has a mean concreteness rating of 4.97 (highly concrete) and 'spirituality has a mean concreteness rating of 1.07 (highly abstract). It just does not seem to make sense to

maintain that concepts of doorknobs are somehow 'conceptually richer' than concepts of spirituality unless we have a clear definition of what constitutes this sort of richness. Indeed, pre-theoretically it might seem more attractive to be able to maintain that the set of representations attached to a concept of spirituality is, at least in some ways, considerably 'richer' than the set of representations attached to the concept of what a doorknob is.

Walker and Hulme (1999, p. 1261) make a similar claim to Romani et al.:

> [concreteness effects] can be best explained in terms of the idea that concrete words benefit from a stronger semantic representation than do abstract words and that the quality or strength of a word's semantic representation contributes directly to how well it can be recalled.

In their discussion section, Walker and Hulme (1999, p. 1267) do spell out what might constitute the 'quality or strength' of a semantic representation. They cite a featured-based model developed by Plaut and Shallice (1993). In this feature-based model, concrete items are 'much richer' than abstract items because they contain 'more consistently accessed features' (Plaut and Shallice, 1993, p. 92). Ultimately, Walker and Hulme's explanation is potentially adequate to the extent that one accepts a feature-based account of meaning, but there are a number of well-known problems with such feature-based accounts (Prinz, 2004). For example, there is the problem of an infinite regress: presumably the features that constitute a concept must themselves be described by sets of features, and although it might be possible in principle to halt this regress, such a mechanism is required before a feature-based account works. More worryingly, it is very difficult to describe a large class of perfectly mundane concepts in terms of features alone: what is the feature list that describes the concept of 'air'? A related issue is that, even if the anti-regress mechanism just mentioned were in place, we do not have to go very far before we find a concept that is difficult to describe in terms of features, and yet is itself a central feature of another concept. For example, one feature of the concept 'dog' might be 'loyal'. We can then ask: what are the *features* of the concept 'loyal'? 'Abstract' concepts such as these are notoriously difficult to characterise in terms of features. This is problematic because we are dangerously close to having no description of the features of an abstract concept *at all*, but at the same time maintaining that concrete items are easier to recall than abstract items because concrete items are associated with 'more features', some of which may turn out to be abstract concepts themselves. There is

a troubling explanatory gap here. Allen and Hulme (2006, p. 79) suggest that 'how well a word is recalled may partly depend on how effectively the semantic representation of that word elicits the appropriate speech output representation at recall'. However, they do not provide an account of what 'effectiveness' means in this context (and perhaps they should not be expected to, given that their focus is on other issues). This is the extent of their account of why these concreteness effects might emerge, although presumably they would at least partially endorse the account proposed in Walker and Hulme (1999). Finally, Miller and Roodenrys (2009) also seem to endorse a feature-based account, and so their explanation is subject to the challenges to feature-based models just discussed.

There is another reason to be cautious about interpreting the concreteness effect in experiment 3. It is very difficult to make numerical predictions in lots of areas of psychology, and these list memory paradigms are no exception. However, I think we should consider how well the results we have obtained match up with theories of why concreteness is psychologically important in the first place. Essentially every theoretical explanation of a concreteness effect involves the idea that there is a fundamental ontological distinction between two categories of cognitive entity. The mental representations that constitute concrete concepts are *different* (richer; more multimodal, more 'grounded', etc.) to the mental representations that constitute abstract concepts. This difference shows up in experimental tasks because of the properties of the human mind-brain. But, if we imagine that these explanations are true, then we might wonder why this fundamental ontological distinction between two kinds of cognitive entity has only produced an effect of *0.3 words*, on average. Perhaps this does not seem problematic. However, it is, arguably, not unreasonable to expect that default processing differences attributed to a profound difference in representational structure should manifest to a greater degree. Furthermore, this kind of theoretical explanation is not really compatible with the fragility of the effect: it is unclear why we obtained marginal evidence in favour of the null hypothesis in experiment 1, if the cause for the effect in experiment 3 is attributed to fundamental structural properties of the human mind-brain. If that were so, then the effect should be much more robust than this, especially given that the experimental design should have maximised its magnitude.

5.7 Summary of Chapter 5

I now summarise the main points of this chapter. List memory experiments have tended to produce concreteness effects such that concrete words are easier to

remember than abstract words. However, it turned out that the ‘abstract’ stimuli featured in these experiments were actually partially made up of midscale words for which the concreteness measure is uninterpretable. There were also other problems with previous list memory experiments, such as low numbers of stimulus controls. This reduces confidence in the reliability of concreteness effects.

In three new experiments, I tried to solve these issues and provide a better test of the existence of concreteness effects. In no experiment was there any evidence for the hypothesis that midscale words are harder to remember than words at the extreme ends of the concreteness scale. The first two experiments returned either null results or marginal evidence in favour of the null hypothesis of no difference between concrete and abstract conditions. The third experiment did produce a small concreteness effect. However, I think that in light of the considerations presented in this chapter and the previous one, we should not take this as incontrovertible evidence that concreteness is a psychologically important variable. The effect was extremely small, and comparable to previous concreteness experiments that featured a large source of noise. This is despite the fact that the design of experiment 3 was such that if the effect was due to concreteness, the effect should have been larger than previous experiments, assuming that concreteness is actually a linear variable. In any case, arguably, the effect was small enough that it is not compatible with the theories that explain it. This is suggestive that something other than concreteness as psycholinguists conceive of it was the reason that one set of words was easier to remember than another set of words. Also, because the midscale problem was present in all previously reported list memory experiments, we really only have one example of a concreteness effect in list memory, which is experiment 3 of this chapter. This is not a large amount of evidence to draw on, and so it seems reasonable to amass more evidence before committing to the idea that concreteness effects in list memory are robust.

Having assessed the prevalence of concreteness effects in list memory, I now turn to concreteness effects in EEG paradigms. Here too we shall see that at first glance concreteness looks like a plausible explanation of the data. However, on closer examination the picture is more complicated, and there are other explanations for various patterns of results that do not require us to appeal to concreteness.

Chapter 6: Concreteness effects in EEG experiments

In Chapter 5, we saw how, even in a standard list memory paradigm, concreteness effects are surprisingly difficult to obtain when the experimental contrast between concrete and abstract stimuli is maximised, and measurement noise is decreased. However, in experiment 3 we did see a small concreteness effect. I now want to consider data from electroencephalography (EEG) experiments. You might think that if the concreteness effect has survived in list memory paradigms, then we can be sure that the midscale disagreement phenomenon is not fatal to experimental concreteness research, and that the issues that this phenomenon raises are nothing to worry about. However, EEG concreteness research is especially interesting for our purposes because it is characterised by some counterintuitive and sometimes contradictory results. I will explain why this is, and then summarise the recent findings in EEG research on concreteness effects. In doing so, I shall point out what I take to be some areas of confusion. Finally, I report a new EEG experiment that measured responses to concrete, midscale, and abstract words in a sentence reading task. Conventional frequentist analysis suggested that there were no statistically significant differences between concrete and abstract conditions. A Bayesian ANOVA suggested that the evidence actually favoured the null hypothesis. I end this chapter with a discussion of the implications of these results, and I offer some reinterpretations of N400 results that have been obtained in previous concreteness experiments (Barber et al., 2013; Holcomb et al., 1999).

6.1 Early EEG concreteness experiments

I will only give a brief overview of the relatively early EEG concreteness work, because as we shall see shortly there are reasons to be cautious about the validity of the stimuli used in those experiments. Holcomb et al. (1999) provide a representative example of the kind of experiment we are interested in. They reported results from two tasks in which participants read sentences that ended either with a concrete word or with an abstract word. In the first task, the final word could also have been either congruent or incongruent with the sentence that preceded it (a 2 x 2 design). Holcomb et al. found that N400s to congruent concrete and abstract words were the same, in

line with work from Schwanenflugel and colleagues (1992; 1983; 1989) in which there were no behavioural differences between concrete and abstract conditions when the stimulus words were embedded in coherent discourse contexts. However, when the sentence-final words were incongruent, N400s to concrete words were larger than N400s to abstract words, even though concrete words were verified quicker than abstract words. One potential explanation of this result was that that the concrete sentences happened to be more anomalous than the abstract sentences because of the nature of the specific meanings of the words, rather than concreteness per se. To check this possibility, Holcomb et al. ran a second task that contained a 'neutral' sentence context which did not bias participants to expect any particular word at the end of the sentence. Interestingly, concrete words still showed higher N400s than abstract words in the neutral condition. Holcomb et al. interpret this finding as showing that the N400 effect in their experiments really is driven by concreteness, as opposed to a confound between how anomalous their sentence endings were across conditions. However, as I have already pointed out, this pattern of results is somewhat confusing: why should a measure that is correlated both with processing difficulty and 'semantic' factors (the N400) produce *lower* amplitudes for abstract words, even though they are verified slower than concrete words?

The reason that we should be cautious about accepting the results of early EEG concreteness studies is because, as Barber et al. (2013) have recently highlighted, all of these previous studies suffered from relatively poor stimulus controls. For example, Holcomb et al. (1999) only controlled for frequency and cloze probability, and nothing else. Since these early studies were conducted, a great many more psycholinguistic variables with behavioural effects have been discovered, such as age of acquisition (AoA) and emotional valence. It might be particularly important to control for AoA because it is now known to have relatively large behavioural effects (Kuperman et al., 2012a) as well as to modulate fMRI and EEG responses (Fiebach et al., 2003; Tainturier et al., 2005). Because abstract words tend to have higher AoAs than concrete words, this could well have contributed to or even obscured differences in the EEG signal between conditions. Note also that the N400 is not the only component that might be affected by any number of lexical variables, and that an impact on one component might give the appearance of lower or higher amplitudes for other components earlier or later in the ERP waveform.

6.2 Barber et al. (2013)

Barber et al. (2013) did control for pretty much every psycholinguistic variable known or suspected to be relevant to single word processing, and so their results are much more robust in this regard. They report a lexical decision task instead of a sentence processing task, in which participants made speeded judgements about the lexicality of concrete and abstract words. Consistent with another lexical decision study in which a higher than average number of lexical variables were controlled (Kousta et al., 2011), Barber et al. found that abstract words were actually recognised *faster* than concrete words. However, the 'normal' N400 differences were still found: concrete words elicited larger N400s than abstract words. At this point, I admit that I find Barber et al.'s report somewhat confusing. You might think that this is a nice result, because now the EEG data and the behavioural data point in the same direction: abstract words were processed faster *and* they elicited lower N400s, and vice versa for concrete words. One potential explanation for this is that it is something to do with the addition of extra stimulus controls in the Barber et al. experiment. However, this does not seem to be the interpretation that Barber et al. (2013, p. 52) come to: they repeatedly state that their results show a 'dissociation' between the behavioural data and the EEG data. This is confusing because what actually seems to have been obtained is evidence consistent with the opposite interpretation: now there is no dissociation between behavioural data and EEG data.

This minor issue aside, there is another potential source of confusion here because some of the previous studies on concreteness effects in EEG paradigms involved sentence processing tasks (Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000), and not lexical decision tasks. In Holcomb et al. (1999), participants read sentences one word at a time and judged whether the sentence as a whole 'made sense'. In West and Holcomb (2000), participants judged whether the final word of a sentence was concrete or abstract. On the other hand, in Kounios and Holcomb (1994), one task was a simple lexical decision task, and another task required participants to judge whether words presented in lists were concrete or abstract. Barber et al. (2013) report a single lexical decision task. Given different task demands and stimulus sets, and the common assumption that the N400 reflects a composite of different neural processes, it is difficult to compare results across these studies. This might not matter too much, because, to the extent that any of the stimulus confounds that Barber et al. tried to rectify had any effects in these previous studies, there is really only one EEG study of concreteness effects with "clean" stimuli, and that is Barber et al.'s lexical decision experiment.

However, the fact that Barber et al. used a lexical decision task as opposed to a sentence reading task raises some questions about their theoretical interpretation of their results. Barber et al. take their results to be incompatible with the 'context-extended dual-coding hypothesis' proposed by Holcomb and colleagues for two reasons. First, Barber et al. obtained a behavioural advantage for abstract words, whereas they argue that the context-extended dual-coding hypothesis predicts a behavioural advantage for concrete words in the lexical decision paradigm. I don't think this argument goes through, because as we saw in Chapter 2, Dual Coding Theory's originator (Paivio, 2013) has repeatedly argued that Dual Coding Theory does not make any behavioural predictions either way in lexical decision tasks. Second, the context-extended dual coding hypothesis holds that the N400 effect in concreteness experiments is due to the higher "semantic richness" of the cognitive resources associated with concrete words. Words that trigger "richer" conceptual representations might prompt more computationally demanding integration or search processes than words that trigger "poorer" conceptual representations, and this might be what the higher N400 for concrete words is indexing. However, since Barber et al. controlled for variables that were supposed to reflect semantic richness and still found the N400 concreteness effect, they argue that this cannot be the case. I'm not sure that this argument goes through either, because, in my view, it's open to question whether the variables that Barber et al. controlled actually do measure 'semantic richness' in the sense in which they seem to be using the term. In reference to previous studies, Barber et al. (2013, p. 48) cite both 'number of features' and 'number of associates' as being measures of semantic richness. But these seem like different and potentially independent measures: it's not hard to think of a word that refers to something that has a small number of "features", but which nevertheless has a relatively high number of first order word associates. For example, it's hard to think of "features of" the word *idea*, but *idea* has non-negligible word association strengths with 17 other words in the Florida Free Association norms (Nelson et al., 2004). By comparison, a straightforwardly feature-ful word like *dog* only has five such associations. And even putting this point aside, in the experiment they report, Barber et al. did not actually control for either of these variables: they state that 'context availability' is the measure of semantic richness they used. A context availability measure is derived by asking participants how hard it is to think of a context in which a word straightforwardly applies. In order for contextual availability to be a measure of semantic richness, then we have to assume that the easier it is to think of such a context, the 'richer' the mental representations triggered by the word are likely to be. I don't know of any evidence either way about whether that assumption is correct.

In any case, Barber et al.'s (2013, p. 51) preferred explanation of their results runs like this: encountering a concrete word triggers 'modality-specific features', by which they mean sensorimotor representations derived from perceptual experience. In contrast, an abstract word '[relies] on emotional associations as well as a variety of other situational and linguistic information'. In a lexical decision task, where words appear in isolation and a participant is told to respond as quickly as possible, 'abstract meanings will therefore receive minimal processing'. This is presumably because in a lexical decision task, there is relatively little 'situational' and/or 'linguistic' information available. Concrete words, on the other hand, will still 'activate and integrate multimodal (sensorimotor) features from distributed cortical networks'. In sum, the EEG concreteness effects are an 'index of meaning activation processes, modulated by the degree of multimodality of the semantic information being integrated (greater for concrete than abstract words)'. The reason that abstract words are responded to quicker than concrete words is because of 'decision and response-selection mechanisms that are extremely sensitive to control'. I now want to try to unpack some of this explanation.

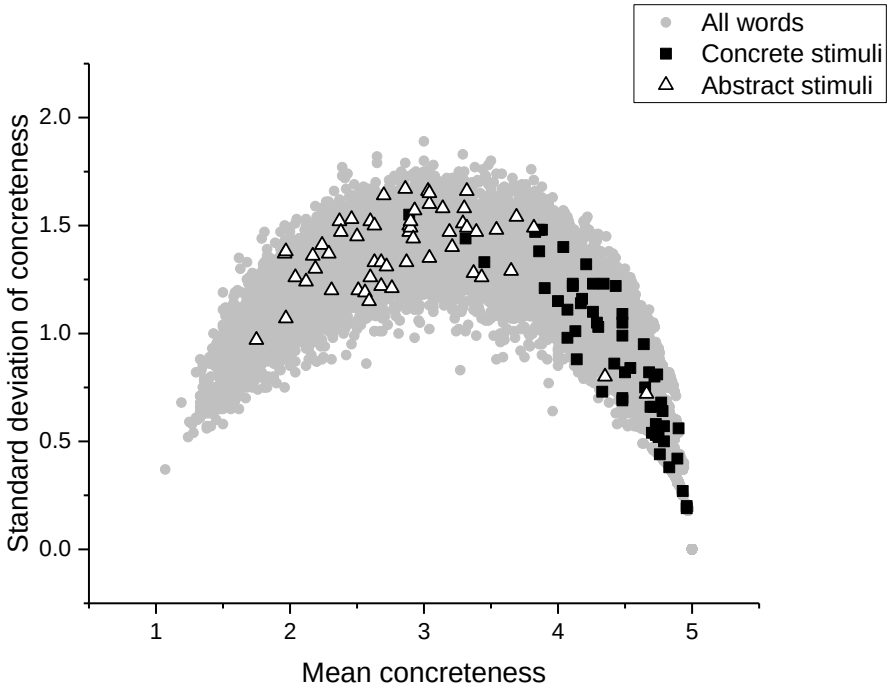
First, Barber et al. explain their behavioural abstractness advantage by appealing to 'decision and response-selection mechanisms', but as far as I can tell they don't say what these are or why they would explain the decrease in response latency for abstract words. The hypothesis might be that, if the language parser/conceptual system detects that fewer multimodal representations have been triggered by a letter string, it can more quickly get to the business of deciding whether that string constitutes a known word. But that hypothesis is *surely* not more compelling than the completely contradictory hypothesis that if the language parser/conceptual system detects that a *high* number of multimodal representations have been triggered by a letter string, it can more quickly get to the business of deciding whether that string constitutes a known word. In fact, you might think that, if the language parser/conceptual system detects that a letter string has triggered a high number of *any* particular sort of representation, then it can be sure relatively quickly that the letter string does constitute a known word, because non-words aren't likely to trigger a high number of mental representations of any sort. So before we accept this explanation, we need some independent evidence in favour of it. In previous work by some of the same authors (Kousta et al., 2011), it was suggested that an advantage for abstract words in lexical decision might be due to the increased amount of 'emotional' information associated with these words over concrete words. However, since emotional valance was controlled in the Barber et al. (2013) study,

that explanation can't work here either: their concrete and abstract words were matched on the emotional valance measure. Finally, Barber et al. (2013, p. 51) suggest that 'abstract words will activate a number of superficial associations with other words' to a greater extent than concrete words. Perhaps the language parser can detect a large number of associative links, activated automatically in response to encountering a string of letters, and it uses this signal to quickly decide that a letter string constitutes a real word. Abstract words have more associative links with other words than concrete words, and this is why they are categorised quicker than concrete words. This would be a plausible enough explanation except for the fact that according to the Florida Word Association Norms, the mean average size of association networks of the abstract words in Barber et al.'s stimuli set were almost exactly the same size as the association networks of the concrete words (13.7 for concrete words; 14.0 for abstract words). So, although it *could* be true in general that abstract words have more associative links than concrete words, this cannot explain the behavioural advantage for the specific abstract stimuli that Barber et al. used. In short, the behavioural abstractness advantage is still something of a mystery.

I think there is also a problem with their explanation of the N400 amplitude differences. According to Barber et al., the N400 indexes 'meaning activation' processes, and these processes are greater in magnitude for concrete stimuli when words are presented in isolation. These meaning activation processes are modulated by how much 'multimodal' information has to be 'integrated'. Concrete words trigger more multimodal information by default than abstract words, and so that's why N400 amplitudes are larger for concrete words. For one thing, it isn't clear that, in order to verify that a letter string is a word of English, any 'multimodal' information about the *referent* of the word should have to be 'integrated' at all. And, even if it did, it's still hard to see how 'integrating' some sensorimotor experience of, say, my nephew, should help my language parser decide that <nephew> constitutes a word of English. However, let's grant that in order to verify that the letter string <nephew> (a concrete stimulus featured in Barber et al.'s experiment) constitutes an English word, some multimodal experience of nephews does have to be 'integrated'. It is *still* unclear why it should be that <nephew> requires *more* multimodal information to 'integrate' than <panic> does, which is an abstract stimulus featured in the same experiment. Other pairs of words chosen more or less at random from the stimulus list also illustrate the point: is it really plausible to suppose that <channel> and <pancreas> (concrete) trigger *more* multimodal information than <dream> and <concert> (abstract)? I am not even sure if I have *any* direct sensorimotor experience of an object I knew to be a

pancreas, but the same is not true of events I identified as being concerts. A potential objection to my line of argument here is that it is a grave mistake to consider individual pairs of stimuli, because of course the N400 concreteness effect is a general aggregate effect that only occurs after large numbers of trials and comes out in the wash after statistical analysis. The effect depends upon aggregate comparisons of responses to groups of stimuli, rather than individual stimuli. My response to this objection is that I do not think that the stimuli I have picked out are an exception. Unfortunately, Barber et al.'s stimuli suffer quite heavily from the midscale disagreement phenomenon, as Figure 6-1 illustrates:

Figure 6-1 Stimuli featured in Barber et al. (2013)



So here, all the conceptual problems relating to the distribution of the concreteness scale that we have seen in previous chapters apply. For a large number of these 'abstract' stimuli, we have no basis for putting them in that category, because approximately half of the participants rated them as being concrete. This makes Barber et al.'s explanation of both the behavioural advantage for abstract words and the N400 amplitude differences less compelling. The theoretical explanation of why these effects are there is inconsistent with the properties that the stimuli which produced the effects seem to have.

So, to summarise: historically it has been found in sentence processing tasks that concrete words are processed faster than abstract words, but that N400s to

concrete words are larger than N400s to abstract words. In a very well-controlled lexical decision study, Barber et al. (2013) found that abstract words were processed *faster* than concrete words, but N400s to concrete words were still larger than N400s to abstract words. However, although the stimuli were well-controlled with respect to lexical variables, they were not well-controlled with respect to the specific statistical properties of the concreteness measure itself. Given that there is still some doubt as to whether the behavioural and EEG data 'match up' in these EEG experiments, I will now describe and report a new sentence processing experiment that maximised the contrast between concrete and abstract conditions in the same way as with experiments 1-3.

6.3 Experiment 4

The new experiment reported here is a single word presentation sentence reading task. Critical trials consisted of a sentence that contained either a concrete word, an abstract word, or a word taken from the middle of the concreteness scale. As with the list memory tasks reported in Chapter 5 the standard deviations of the concreteness ratings of the concrete and abstract words were tightly controlled and minimised, whereas the standard deviations of the concreteness ratings of the midscale stimuli were maximised. Of interest were the N400 amplitudes to concrete and abstract words that were taken from the extreme ends of the scale, and those words about which participants disagreed. The midscale condition was included in order to check the possibility that EEG results may diverge from the behavioural results obtained in Chapter 5, where no behavioural difference between concrete and midscale conditions was obtained in list memory tasks. This possibility is worth checking precisely because, as we saw in the introductory sections of this chapter, EEG concreteness research is marked by the finding that the EEG data and the behavioural data do not match up in the way that we might expect.

A sentence processing task was chosen over a lexical decision task for a number of reasons. If the N400 concreteness effect is driven by a need (or automatic tendency) to 'integrate' information (Barber et al., 2013), then we would expect more information to be integrated in a task in which a participant generates an interpretation of a sentence, than in a task in which all they do is verify whether unconnected letter strings are words or not, as is the case with a lexical decision task. Participants were instructed to read sentences presented one word at a time, and judge whether the sentence contained any errors. Half of the sentences in the experiment contained an anomalous word that obviously did not fit well with the rest of the sentence. This

ensured that participants were motivated to attend to every word in each sentence. A novel aspect of this new experiment was that the experimental manipulation itself did not involve any anomalous words. Instead, the target concrete, midscale, and abstract words appeared near the beginning of sentences, before it was possible to identify whether the sentence as a whole would be anomalous or not. This modification to the normal paradigm (Holcomb et al., 1999; Kutas and Hillyard, 1984), in which responses are measured to words at the end of a sentence, was also made for numerous reasons. It has been shown repeatedly that, if enough supporting context is provided, EEG and behavioural responses to concrete and abstract stimuli are similar (Holcomb et al., 1999; Schwanenflugel and Shoben, 1983). However, if the concrete/abstract words appear at the beginning of a sentence, then when the target word is encountered, a participant has no basis on which to expect anything in particular, because not enough context has been provided in order to support or predict any interpretation. In that sense, from the point of view of the language parser, the situation is somewhat similar to that found in a lexical decision task, when a word is presented in isolation. However, because the participant ultimately has to make a message-level decision that requires interpreting the whole sentence, they are still encouraged to process each word more fully than they would have to in order to complete a lexical decision task. Therefore, the chances of finding an N400 concreteness effect should be maximised *if*, as Barber et al. (2013) suggest, this effect is driven by the integration of information triggered upon encountering a word.

This set-up provides the additional advantage that the cloze probability of the target concrete, abstract, and midscale words is essentially 0 in every case. Each target sentence started with a neutral frame (such as a definite article or possessive) that made it impossible to predict what word would follow (see stimuli section for details). This provides an easy way of controlling for cloze probability, which is especially important because of the large effect it exerts on N400 amplitudes (Kutas and Hillyard, 1984). One final advantage is that this experimental format allows us to analyse responses to all target words without worrying whether the sentence as a whole is anomalous or not, or indeed whether the participant made a correct decision about the validity of the sentence. That is because the anomalies all appeared after the initial part of the sentence, so that, from the parser's point of view, words near the beginning of the sentence should not be considered anomalous, and so N400 responses to them should be relatively 'natural' N400s that reflect normal reading and interpretation processes.

Participants

Initially, 24 native English speakers with no reported neurological disorders were recruited from the University College London SONA subject pool, of which 21 turned up for their session. Of these, two participants were excluded from analysis because their EEG data contained too many artefacts and/or too much noise. Two more participants were excluded because their response accuracies for detecting anomalies were extremely low (58% and 65%). 15 participants were male; mean age = 26.8 (9.9). The number of participants included in the analysis was therefore 17. Each participant was paid £20 for their time.

Procedure

The experiment took place in an air-conditioned, sound-proof, electrically shielded booth specially constructed for the collection of EEG data. Participants read sentences presented one word at a time in the centre of a computer screen. Each word was presented for 300ms, and was then replaced by a blank screen lasting 250ms. Then, the next word in the sentence was presented. The SOA was therefore 550ms, which is a large enough window to examine the N400 effect of interest. At the end of each sentence, an on-screen prompt reminded participants to indicate whether the sentence contained any errors. Participants used the F and J keys on a standard computer keyboard; yes/no keys were counterbalanced across participants. While the sentences were being presented, the participant was instructed to try to keep movements and blinking to a minimum. At the end of each sentence, the participant could move and blink if they wanted to. The experiment was broken up into 6 blocks of sentences. After each block, the participant was prompted to rest for as long as they liked before continuing on the next block. During the rest breaks, the experimenter entered the booth to check on the participant and make adjustments to any electrodes that appeared to be noisy, or suffered from high impedence levels. Noise levels and impedence for every electrode were monitored constantly throughout the experiment, and if the impedence of any electrode rose above 25mV, the experiment was paused so that the impedence could be lowered.

EEG recording and analyses

EEG was recorded from 32 scalp electrodes using the 20-10 system. There were four external electrodes. Two electrodes were used in order to detect trials contaminated by blinks and other eye movements (one below the left eye; one next to the right eye). Two electrodes were placed on the left and right mastoids; the average of these mastoid electrodes provided the reference measure. EEG was digitised at a rate of 512Hz. As per the recommendations of Luck (2014), a gentle bandpass filter was

applied offline (0.01-128Hz). EEG was epoched in 750ms segments, each consisting of a 200ms pre-stimulus baseline period and a 550ms post-stimulus period. The specific window of analysis for the N400 effect was 300-550ms post stimulus. This window was chosen because it overlapped maximally with the N400 windows chosen in other EEG concreteness experiments (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000), and independently of the obtained mean average amplitudes. Trials containing blinks and other artefacts were marked and removed from the averaging process (83% of trials remained, distributed equally across conditions).

Statistical analyses were based on the repeated measures ANOVAs employed by Barber et al. (2013). The main analysis was a 3 x 2 x 2 factorial design in which there were three levels of CONDITION (Concrete versus Midscale versus Abstract), two levels of LATERALTY (left versus right), and two levels of ROSTERALITY (front versus back). Midline and lateral line electrodes were omitted in order to partition the scalp into four quadrants, each consisting of six electrodes. The dependent measure was the mean amplitude between 300 and 550ms after the presentation of the stimulus word. The three terms of interest are: the main effect of condition, the interaction between condition and laterality, and the interaction between condition and rosterality. Note that this design is somewhat unusual in that there is no behavioural measure relevant to the concreteness effect of interest. Although some sentences were anomalous, and participants were asked to detect these anomalies, the point in the sentence at which the target words occurred was too early for there to be any evidence of an anomaly. Anomalous sentences were only included in order to provide a way of checking that participants were attending to the sentences.

Stimuli

There were three conditions in the experiment: concrete, abstract, and midscale. Each condition consisted of 66 sentences, and each sentence contained a target word that was concrete, abstract, or taken from the middle of the concreteness scale. Sentences were constructed according to the following constraints. The target word appeared near the beginning of the sentence in subject position. Within a condition, 43 sentences started with the sequence “The [target word]…”, 9 started with “The man’s/woman’s [target word]…”, and 14 started with a name in the genitive (“David’s [target word]…”). Full example sentences and patterns are given below. For the first

three example sentences, the target words are in underlined bold font. For the final example sentence, the anomalous word is in underlined bold font.

Concrete example: The man's **aftershave** was cheap.

Midscale example: Robin's **degree** was in chemical engineering.

Abstract example: The **oversight** was costing the company lots of money.

Anomaly example: The vigilante was **skipping** from the law.

The small number of different starting sequences was used in order to make it easier to generate natural-sounding sentences for each target word. The words that immediately followed the targets were also controlled. 46 sentences in each condition continued with the past tense of the verb 'to be': "The [target word] was…". 22 sentences in each condition continued with a prepositional phrase: "The [target word] of…". These constraints were introduced in order to make sure that the ERPs to the words following the target words were all generated in response to the same stimuli, so that the impact of subsequent words on the ERP waveforms would be the same across conditions. Half of the sentences in the experiment contained a single word that created a semantic anomaly. The position of this anomalous word varied from sentence medial to sentence final positions so that participants could not predict where it would occur, and would therefore be motivated to read the whole of every sentence in order to detect anomalies. The experiment also contained filler trials of varying syntactic structure; half of these fillers also contained anomalies. Participants would therefore, hopefully, not notice any of the relatively small number of syntactic patterns that made up the sentences of the critical trials.

The target words in each condition were controlled for the following psycholinguistic variables: word frequency, age of acquisition, emotional valance, mean bigram frequency, number of syllables, number of morphemes, and length in letters. Cloze probability was also controlled by virtue of the fact that target words appeared near the beginning of sentences, after neutral sentence initial frames. The mean concreteness scores of the concrete words were all 4 and above, the mean concreteness scores of the abstract words were all 2 and below, and the scores of the midscale words were all between 2.5 and 3.5. The standard deviations of the concreteness scores of the concrete and abstract words were all below 1. The standard deviations of the concreteness scores of the midscale words were all above 1.4. Variable information was amalgamated from: the Brysbaert et al. (2013)

concreteness norms, the Warriner et al. (2013) emotion norms, the Kuperman et al. (2012a) age of acquisition norms, and the English Lexicon Project (Balota et al., 2007). The stimuli across conditions were controlled on these variables by using the MATCH algorithm (Van Casteren and Davis, 2007). Table 6-1 displays the psycholinguistic variable information for the stimuli featured in each condition.

Table 6-1 Properties of stimuli featured in experiment 4

Condition	conc_m	conc_sd	AoA	freq	nsyll	length	bg_m	valence	nmorph
Concrete	4.6	0.8	10	3.5	2.4	7.7	3639	1.1	1.8
Midscale	3	1.5	10.2	3.6	2.5	7.5	3770	1.1	1.8
Abstract	1.6	0.9	10.2	3.5	2.5	7.7	3643	1.1	1.8

Table legend:

conc_m	mean concreteness
conc_sd	mean standard deviation of concreteness
AoA	mean age of acquisition
freq	mean Zipf frequency
nsyll	mean number of syllables
length	mean length in letters
bg_m	mean bigram frequency
valence	mean emotional valence score (0 = neutral, 4 = highly valanced)
nmorph	mean number of morphemes

Results

Response accuracies for each condition were:

Concrete: 87.9%
Midscale: 85.3%
Abstract: 83.1%

These accuracy rates indicate that participants were attending to the task, although they made slightly more mistakes with sentences which began with abstract subjects than sentences which began with concrete subjects.

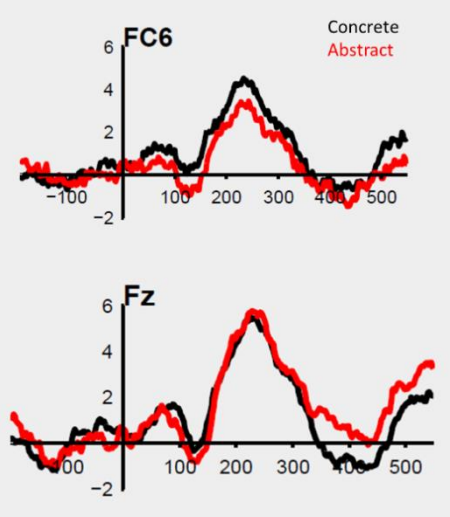
In a 3 x 2 x 2 repeated measures ANOVA, the main effect of condition was not statistically significant ($F(2, 16) = 0.686$, $p = 0.511$, partial $\eta^2 = 0.041$), the interaction between condition and rosterality was not statistically significant ($F(2, 16) = 0.038$, $p = 0.963$, partial $\eta^2 = 0.002$), but the interaction between condition and laterality *was* statistically significant ($F(2,16) = 3.338$, $p = 0.048$, partial $\eta^2 = 0.173$). Table 6-2 below displays the mean amplitudes obtained between 300-550ms by condition and laterality.

Table 6-2 Mean amplitudes in mV between 300-550ms by laterality

	Left	Right
Concrete	0.1	0.1
Midscale	0.44	0.2
Abstract	0.75	0.22

Looking at table 6-2, the figures suggest that the interaction was driven by the fact that for concrete words, N400 amplitudes recorded on the left and right hemispheres were almost identical, whereas for abstract words, N400 amplitudes were lower on the left hemisphere than on the right (recall that the lower an N400 amplitude, the higher the mean voltage will be, because it is a negative-going deflection in voltage). In order to investigate the interaction, two post-hoc repeated measures ANOVAS were run. These analyses compared N400 amplitudes for concrete versus abstract words in the left hemisphere and the right hemisphere respectively. The left hemisphere ANOVA did not produce a statistically significant effect of concreteness: $F(1, 16) = 2.373$, $p = 0.143$, partial $\eta^2 = 0.129$. The right hemisphere ANOVA did not produce a statistically significant effect of concreteness: $F(1, 16) = 0.105$, $p = 0.75$, partial $\eta^2 = 0.007$. Grand average waveforms for concrete versus abstract words at two electrode sites are displayed in figure 6-3 below. These results do not provide evidence for a main effect of concreteness such that N400s to concrete words are larger than N400s to abstract words, and nor do they provide evidence that N400s to midscale words are different to N400s to words from the extreme ends of the concreteness scale.

Figure 6-2 Grand average waveforms for concrete and abstract words at 2 electrode sites



FC6 and Fz were chosen for two reasons. They are approximately the same scalp locations as ones Barber et al. (2013) graphed, and at Fz the difference between conditions is largest. Negative voltages are plotted down. The midscale waveform has been omitted in order to make the graphs clearer.

Of course, a p-value above 0.05 on its own cannot be taken as evidence in favour of a null effect, and, although the small number of participants featured in the present experiment (n = 17) is typical of EEG studies, we might also worry that the null results here are driven by a lack of statistical power. Therefore, as in Chapter 5, I will also present some simple Bayesian analyses designed to address these potential issues. If the Bayes Factors produced by a model comparison analysis of the data are between 1/3 and 3, then we have evidence *for* the experiment being inconclusive. However, if the Bayes Factors of this model comparison analysis are below 1/3, then we have evidence in favour of the null hypothesis, that is, we have evidence for there being no main effect of concreteness or interaction between concreteness and electrode location.

The same data with the same 3 x 2 x 2 factorial design were entered into a Bayesian Repeated Measures ANOVA using JASP (JASP Team, 2018), using the default settings and priors, examining effects across matched models. This analysis compares the likelihood of all models that contain a particular term with the likelihood of otherwise equivalent models from which that term has been removed. We are particularly interested in the likelihood of models that contain a condition term (concrete, midscale, abstract), versus the likelihood of models that do not include that term. The Bayes Factor for the inclusion of the condition term (conceptually similar to assessing the main effect of condition) was 0.14. This indicates that models which include a condition term are 0.14 times as likely as models which do not include the term, given the data and the models. Or, equivalently, we could say that models which do not include the term are approximately 7 times more likely than models which do include the term. The Bayes Factor for the inclusion of the condition*rosterality interaction was 0.09. The Bayes Factor for the inclusion of the condition*laterality interaction was 0.15. Since these Bayes Factors are all below 1/3, this analysis suggests that instead of simply not finding evidence in favour of a concreteness effect, we have actually found evidence *for* an absence of this effect. Table 6-3 below displays the full results of the matched models analysis.

Table 6-3 Summary of Bayesian ANOVA for experiment 4

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$
concreteness	0.263	0.125	0.146
laterality	0.263	0.225	0.312
rosterality	0.263	0.94	2.021e +17

concreteness ∗ laterality	0.263	0.005	0.138
concreteness ∗ rosterality	0.263	0.012	0.092
laterality ∗ rosterality	0.263	0.048	0.211
concreteness ∗ laterality ∗ rosterality	0.053	1.493e -5	0.199

Interestingly, this is one instance in which the output of frequentist and Bayesian analyses is different. The frequentist analysis suggests that the condition*laterality interaction is statistically significant, whereas the Bayesian analysis suggests that we should prefer models of the data that do not include a term for that interaction. Given the lack of a main effect of concreteness in both the omnibus ANOVA and the post-hoc ANOVAS; the fact that the interaction was marginal; and the fact that other studies have not found a condition*laterality interaction, I am inclined to prefer the more conservative Bayesian interpretation of the data.

6.4 General discussion

The results presented here contrast with those presented in other concreteness EEG research (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000). All of these studies found a statistically significant main effect of concreteness on the amplitude of the N400, whereas the current experiment did not. However, it has been found previously that concreteness effects in EEG paradigms do disappear under certain conditions. As noted above, Holcomb et al. (1999) found that when concrete and abstract words ended sentences in a congruent and expected way, the N400s to each type of word were the same. One potential reason for the lack of a main effect in the experiment presented here is that perhaps the concrete and abstract target words appeared in congruent sentence frames, and so that explains the lack of a main effect. I don't think this explanation is valid for two reasons. Firstly, the target words in the current experiment appeared near the beginning of the sentences, and were preceded only by bare definite articles or generic possessives. So the participants had no basis on which to expect any particular word at this point in a sentence. There was therefore no supporting context of the kind that eliminated the concreteness effect in Holcomb et al.'s experiment. Secondly, Holcomb et al. demonstrated that N400s to concrete words were still larger than N400s to abstract words even in neutral sentence contexts, and so *if*

concreteness was the reason that N400 differences were obtained, I think we would also expect an N400 difference in the current experiment, where the beginnings of each sentence were also neutral with respect to which word would follow. Furthermore, Van Petten and Kutas (1990) found that lower frequency words elicited larger N400s than higher frequency words, but only when they were presented near the beginning of sentences, and not when the words appeared at the end of sentences. Therefore, we have at least one other experiment which suggests that N400 differences occur at the beginning of sentences to a greater extent than at the end. If concreteness does drive a difference in N400 amplitude, then we should therefore expect to see this difference for target words near the beginning of sentences.

The results obtained in the present experiment suggest that concreteness may not be the reason for N400 amplitude differences, after all. If that is the case, then we need explanations for why it is that other researchers have found EEG concreteness effects in other experiments. I will offer two such explanations for the studies we have considered in most detail: Holcomb et al. (1999) and Barber et al. (2013). One alternative reason for why Holcomb et al. (1999) found a difference in N400 amplitude in their neutral condition was that, although the cloze probabilities of the concrete and abstract target words in their neutral sentences might have been closer than in the anomalous sentences, these cloze probabilities still might have been unequal in absolute terms. The example neutral sentences they give are:

"It happened because of her **mood**."
"Larry said it must have been the **wine**."

"Wine" is a target concrete word, whereas "mood" is a target abstract word. Note that although these sentence frames are *relatively* neutral, it still is quite unlikely that the true cloze probabilities of the sentence-final words are the same across conditions. To see why this might be, consider the following sentences:

"It happened because of her *wine*"
"Larry said it must have been the *mood*"

It seems to me that these latter two sentences are less natural-sounding than the first two sentences, but all we have done is swapped the sentence-final target words. If these sentences really were neutral with respect to expectations about the final word, as Holcomb et al. hoped, then this just wouldn't be the case. For this reason, it's possible that the cloze probabilities of the target words in the experiment I report in this chapter were more similar (and smaller) than in the experiments reported in

Holcomb et al., because the target words appeared near the beginning of sentences. This suggests that in Holcomb et al.'s experiment, the N400 differences that have been found so far between concrete and abstract words might not be due to concreteness, but instead might still be due to a confound between the cloze probabilities of the concrete words versus the abstract words. And indeed, as Barber et al. (2013) highlight, it could also be due to any number of other psycholinguistic variables that tend to be correlated with the concreteness measure (e.g., AoA) which Holcomb et al. did not control for.

As we saw above, Barber et al.'s preferred explanation of their lexical decision EEG results is that concrete words trigger more multimodal information by default than abstract words. So, if that's true, it is unclear why it would be that the concreteness N400 effect does not occur for words near the beginning of sentences, where presumably at least as much multimodal information is triggered as in a lexical decision task. However, it is still the case that Barber et al. (2013) found a difference between the N400 amplitudes to concrete and abstract words in their lexical decision task, and if we entertain the idea that concreteness is not really the driving factor behind differences in N400 amplitudes obtained in other experiments, then we also need an explanation of this result. Here is a potential explanation. The amplitude of the N400 is correlated with how expected a word is (Kutas and Hillyard, 1984). Consider Table 6-3 below, which displays all of the stimulus words featured in Barber et al.'s experiment.

Table 6-4 Words featured in Barber et al.'s experiment

Concrete				Abstract			
gig	rod	Voice	corridor	mood	maternity	horror	protest
cue	nephew	Tribe	liquor	amour	apology	marriage	reflection
disease	column	product	prong	woe	quest	flutter	number
guest	monsoon	channel	cable	delirium	fury	temper	slumber
ounce	rack	freight	machinery	fun	sum	bargain	demon
cancer	block	author	sound	pleasure	magic	luxury	peep
ether	drape	material	weapon	marvel	joke	happiness	warmth
date	synagogue	starch	plate	triumph	spree	dream	fashion
duke	manure	palette	garment	romance	beauty	burden	space
bureau	university	widow	medicine	confusion	crime	wealth	sneer
creature	pocket	troop	pancreas	haste	plunge	danger	whiteness
relic	warehouse	clove	equipment	frenzy	panic	expanse	angel
air	lobby	jack	insect	joy	minute	grief	concert
estate	cartilage	jersey	stomach	agony	midnight	love	whoop
mountain	thong	thyme	belt	thrill	ghost	paradise	dozen

The concrete words in Barber et al.'s stimulus list nearly all refer to conceptually unrelated medium sized objects. Where there are connections between the referents of the concrete words, they tend to be only between pairs (e.g., *cable-machinery*, *cancer-disease*). Although individual intuitions about the abstract stimuli will vary, it's still clearly the case that, in contrast to the concrete stimuli, a large number of them form a semantically coherent group that we might characterise roughly as 'physical or emotional human states' (highlighted in grey in Table 6-3). By my count, there are 25 of these in the abstract condition, which is nearly half of the 60 stimuli that made up the whole condition. That means that, of the real words of English that participants saw over the course of the experiment, approximately one quarter were potentially semantically related to each other, and these were all part of the abstract condition. Although the stimuli were presented in a random order and interspersed with concrete words and non-words, repeated exposure to exemplars of this set may have primed other members of this set, which would make them more expected than the rest of the stimuli in the experiment. This could be why the N400s to the abstract words in Barber et al.'s experiment were lower than the concrete words: it is well known that expected stimuli prompt smaller N400s than unexpected stimuli. Indeed, this was one of the original considerations when hypotheses about the cognitive processes underlying the N400 were being formed (Kutas and Hillyard, 1984).

To put the point another way: suppose that I ran a lexical decision experiment in which 25% of the real-word stimuli referred to household tools (hammer, spade, trowel, drill…), and the other 75% were randomly drawn from truly unrelated domains. My guess is that, on average, N400s to these tool words would be lower than to the unconnected non-tool words, but this would have nothing to do with any particular properties of the tools or the words we use to refer to them. This explanation is admittedly speculative, but I think that it is at least as plausible as the explanation that we would have to entertain if we accepted that the N400 amplitude differences really were driven by concreteness (that is, the explanation that requires that it be the case that encountering the letter string <pancreas> triggers more sensorimotor information than encountering the letter string <concert>). Furthermore, decreased N400 amplitudes for semantically primed words in lexical decision tasks have been found in other experiments (Bentin et al., 1985; Holcomb, 1993), and so it would not be surprising to observe them in Barber et al.'s experiment. This alternative explanation also has the benefit that it appeals to what is already known and (relatively) uncontroversial about the amplitude of the N400: it decreases as a function of how expected a stimulus is. Of course, this alternative explanation also might equally apply

to the behavioural advantage Barber et al. found for abstract words. A large number of the words in their abstract condition could potentially prime other abstract words in that condition, and this could plausibly decrease decision latency for those words.

6.5 Summary of Chapter 6

To conclude: in this chapter, we considered evidence that N400 amplitudes to concrete words are larger than N400 amplitudes to abstract words, as has been found in a number of previous studies (Barber et al., 2013; Holcomb et al., 1999; Kounios and Holcomb, 1994; West and Holcomb, 2000). This has been the typical finding in EEG concreteness research, despite the behavioural evidence that concrete words are easier to process than abstract words. These two findings are in tension because N400 amplitudes are typically larger for stimuli that are harder to process than stimuli that are easier to process. However, Barber et al. (2013) showed that earlier concreteness EEG studies potentially suffered from a large number of stimulus confounds. In their well-controlled lexical decision task, they still obtained larger N400 amplitudes for concrete words, but they found a behavioural advantage for abstract words. I argue that their theoretical explanations of these results are not consistent with the properties of the specific stimuli that featured in their experiment. They claim that concrete words trigger more 'multimodal information' than abstract words, which explains the larger N400 for concrete words, whereas abstract words are more susceptible to 'decision mechanisms', which explains the behavioural advantage for abstract words. However, absent any evidence, it is not plausible to suppose that the concrete stimuli featured in their experiment actually do trigger more 'multimodal information' than the abstract stimuli. I also suggest that it is very hard to think of a 'decision mechanism' that would actually explain the behavioural abstractness advantage. In any case, the stimuli featured in Barber et al.'s abstract condition suffered from the midscale disagreement phenomenon we have examined in previous chapters.

In order to address this problem, I reported the results of a new sentence processing experiment that was designed to maximise the chance of finding an N400 concreteness effect, *assuming that Barber et al.'s hypothesis was correct.* Not only did this experiment not find evidence for a statistically significant difference between N400s to concrete and abstract words, Bayesian analyses suggest that we should actually prefer the null hypothesis that there is no such difference. I have offered some explanations for why it is that previous studies found an N400 concreteness effect, whereas the current one did not. Although Holcomb et al. (1999) tried to control for

cloze probability in the neutral condition of their experiment 2, it is unclear whether their neutral sentences really did create equal cloze probabilities across conditions. And Barber et al. are right to point out that other stimulus confounds may well have affected the results of all of Holcomb et al.'s experiments. Although Barber et al. (2013) controlled for an impressive array of psycholinguistic variables, nearly half of the words that made up their abstract condition could have potentially primed each other. These issues could potentially have contributed to a lower N400 amplitude to abstract stimuli. This point, coupled with the fact that evidence *against* an N400 difference between concrete and abstract stimuli was obtained in the experiment reported here, suggests that evidence for the N400 concreteness effect in general is not as strong as is generally believed. Although this might seem like a potentially unwelcome conclusion, I think that, if the analyses I have presented in this chapter are along the right lines, then actually we are in a much better position than we were in before regarding concreteness EEG research. That is because, as McRae and Jones (2013) note, a satisfying explanation of the alleged N400 concreteness effect has been someone elusive. If I am right, then this explanation has been elusive because these effects were not due to concreteness, after all. The upshot of all this is that we have more evidence against the general validity and utility of the concreteness measure in psycholinguistic investigations of the conceptual system. In the next chapter, I shall summarise what we have seen in Chapters 4-6, and draw together my response to objection 1.

Chapter 7: Response to objection 1 (concreteness effects are fragile)

In Chapter 3, I argued that we do not possess a concept (a unitary cognitive resource) for every word we know. Using the alleged abstract concept JUSTICE as an example, I tried to convince you that positing this concept comes with lots of costs but no theoretical benefit. I suggested that we can get by without positing JUSTICE, and that explanations of human behaviour and cognitive processes that we communicate about by using the *word* 'justice' do not contain JUSTICE. I also tried to show that this position comes with a substantial benefit: theories of concepts are in a much better shape if they do not have to deal with the problems that JUSTICE causes them. I also suggested that this strategy could be used to rule out other potentially problematic 'abstract' concepts. We considered two important objections to my arguments. The first objection was:

> If some proposed abstract concepts don't actually exist, then how can it be that reliable experimental effects are obtained by measuring responses to 'concrete' stimuli and comparing them to responses to 'abstract' stimuli? Surely this empirical evidence indicates that there must be something neuro-psychologically real and theoretically principled about the concrete-abstract distinction, and that there is a reliable relationship between words and concepts.

The response I want to make to this objection will probably seem obvious by now. In Chapters 4-6, I presented evidence and analyses to the effect that experimental concreteness effects are just not as reliable as is generally assumed. In Chapter 4, we saw that the way that concreteness is operationalised in psycholinguistic experiments is hugely problematic. At the concrete end of the scale,

perhaps the measure functions more or less adequately. But at the abstract end and in the middle of the scale, the validity of concreteness norms is highly dubious. If we want to construct psycholinguistic experiments featuring 'truly' abstract words that participants are highly likely to be familiar with, then we end up with a pool of less than 300 items to draw on. And, as we saw in Chapter 4, a large number of these are very morphologically complex and/or idiomatic. If there *is* a reliable relationship between words and concepts, and concreteness is supposed to index a psychologically instantiated difference between two kinds of mental resource, then it is very strange that there are fewer than 300 words of English that reliably pick out one of these kinds. In the middle of the concreteness scale, the problem is worse. Taking the mean value of diametrically opposing judgements gives the illusion that participants treat concreteness as a scale, but in fact they do not. For nearly every word in the middle of the concreteness scale, mean values do not reflect participants' judgements. Even more worryingly, in every stimulus list that I have looked at, the 'abstract' stimuli used in experiments were not actually abstract. Instead, they were simply those words about which participants tended to disagree. This undermines the reliability of concreteness effects because we just have not been comparing responses to concrete stimuli with responses to abstract stimuli.

In chapters 5 and 6, I presented some new experiments that were designed to assess the implications of these issues with the concreteness scale. I tried to both maximise the experimental contrast between concrete and abstract items, and significantly reduce the presence of noise created by variability in ratings. In all experiments, the conditions were such that both the chances of finding a concreteness effect and the magnitude of that effect should have been maximised. Despite this, in 3 out of 4 experiments, I obtained evidence for the null hypothesis. I also pointed out that, even if we had obtained evidence for concreteness effects, the theoretical explanations of why these effects occur all fail in various ways. Typically, researchers assume that words with lower concreteness ratings trigger less 'multimodal' information than words with higher ratings, but, if we actually look at these words that feature in experiments, we see that this simply isn't true. There is no reason to believe that the words featured in the abstract conditions (e.g. 'concert') in experiments do trigger less multimodal information than the words featured in the concrete conditions (e.g. 'pancreas'). In Chapter 6, I also suggested that there may be other reasons for why researchers appear to have obtained evidence for concreteness effects that do not appeal to concreteness, such as priming effects and differences in cloze probability across experimental conditions.

I take all of this to show that concreteness is just not as reliable a psycholinguistic variable as it is believed to be. Note that I am not arguing that there definitely are no such things as concreteness effects. I have simply presented evidence that they are much more elusive than we would expect them to be, if the theories that explain them are true. Furthermore, if there are such things as concreteness effects, then the way we operationalise the variable now is a flawed way of trying to find them, for the reasons detailed in Chapter 3. I think that if we put all of this together, then we are left with the conclusion that evidence for concreteness effects is not very strong. Explanations of concreteness effects often require that words and concepts are in a reliable correspondence with one another. I am trying to convince you that words and concepts are *not* in a reliable correspondence with one another. In order to do this, I have presented evidence that concreteness effects are not actually reliable experimental outcomes. We need more evidence in order to determine the status of concreteness effects. That being so, I hope this provides a way of mitigating objection 1, at least for moment. I argued in Chapter 3 that at least some words of English do not pick out elements of a theory of concepts. If there were strong evidence that the concreteness scale indexes reliable processing differences between concrete and abstract words, then this would be problematic for my argument. This is because the concreteness scale assumes that there is a reliable correspondence between words and concepts. However, we *don't* have to assume that words and concepts stand in a reliable correspondence with one another in order to explain concreteness effects, because the experiments I have reported here suggest that that concreteness effects do not obtain in a way that licences this assumption. I will now move on to provide a response to objection 2.

Chapter 8: Response to objection 2 (there is more to meaning than concepts)

Recall from Chapter 3 the second objection to my arguments about the relationship between words and concepts:

> If there is no unitary cognitive resource that corresponds to the word 'justice', then what is the meaning of the word 'justice' – how do we understand each other when we use the word 'justice', or other abstract words for that matter?

My response to objection 2 will take the form of a sketch of a theory of communicative success that does not require that words encode concepts. I argue that Relevance Theory, a popular theory of communication, works better if it relaxes certain assumptions about word meanings and concepts. It's very often assumed that 'the meaning' of a word is a theoretically important notion, and that this meaning is a concept. As we saw in chapters 2 and 3, all of the psycholinguistic research on the concrete-abstract distinction assumes that word meanings and concepts are the very same things, and philosophical theories of thought also assume that words encode concepts. Instead of providing a knock-down argument that this is not the case, I wish to show that with regard to explaining how communication occurs, there is an alternative view that is just as viable. This alternative view denies that meanings and concepts are the same things, and it emphasises that it is more useful to think of meaning as being a property of an interpretation of an utterance, as opposed to a property that a word 'has'. If this alternative view really is just as viable as the conceptual view (the view that word meanings just are concepts), then the default assumption that word meanings just are concepts is not justified. Instead, it would have to be shown that the conceptual view has advantages that the alternative does not. I hope that this deflates objection 2, because with this alternative view in hand, we would have a story to tell about the word 'justice' and meaning.

8.1 JUSTICE, 'justice', and meaning

Consider the question: "what is the meaning of the word 'justice'?". We might be asking, "what cognitive resources should we attribute to Smith in order to explain how he generated a felicitous interpretation of an utterance containing the word 'justice'?". *Or*, we might be asking, "if we polled every member of a language community, how would they respond to the question of what the meaning of the word 'justice' is?". I think it is clear that these two questions are about different things. The first question is a question that a certain kind of cognitive theory should answer. The second question is a question that a certain kind of sociolinguistic investigation could answer. I take it that the philosophers and psychologists we discussed in Chapter 3 are generally interested in the first question when they talk of meanings and concepts, although they do not necessarily couch it in those terms.

So, what resources *would* we need to attribute to Smith in order to explain his generating a conversationally felicitous interpretation of an utterance containing the word 'justice'? What is there, "in the mind", that corresponds to the word 'justice'? The standard answer is that we need to attribute to Smith a concept, JUSTICE. I have spent much of this thesis exploring various problems that the notion of JUSTICE raises. So here is my alternative suggestion. Smith needs some grammatical knowledge of English (which everybody would accept), the ability to draw inferences (which is the cornerstone of many accounts of utterance interpretation), and a stored set of episodes of past encounters with/usages of[9] the word 'justice'. I should note that this is far from being a novel idea (see, for example, Recanati's (2004) brief discussion of what he calls 'meaning eliminativism' and Allott and Textor (2017)), although I hope to draw out some new issues in a particular way. Smith can attempt to understand what his interlocutor meant by their use of an utterance containing the word 'justice', by drawing inferences on the basis of the specifics of the context of the conversation, and the information instantiated in one or more of his past usage episodes. But he doesn't necessarily need a JUSTICE concept.

An obvious concern here is that we need some explanation of what a context is, and how Smith draws appropriate inferences with respect to it. But there are already powerful theories of language interpretation and communication that make detailed suggestions about how we draw inferences with respect to elements of the discourse context and what principles govern these inferences. Relevance Theory is

[9] I will use 'past usage' as a blanket term to refer to both of these options

one prominent example (Sperber and Wilson, 1995). These theories tend to be committed to the idea that words encode concepts, and, at first glance, they might look incompatible with my account for this reason. But I don't think this is *necessarily* true: the mechanics of, say, Relevance Theory (accessibility of encyclopaedic information, contextual assumptions/implications, the presumption of Optimal Relevance, chains of inference) might just as well operate on information contained in stored episodes of past usages of words, as on information stored with concepts encoded by words. So I am not suggesting that Smith understands someone's use of the word 'justice' *just* by consulting an appropriate stored episode of a past usage: I think I can avail myself of the kinds of capabilities that feature in other accounts of language interpretation.

I will attempt to give a rough outline of how the alternative view might look on a Relevance Theory-style account. I will not be able to provide a fully fleshed-out story. I just want to motivate the idea that at least it's plausible that Relevance Theory (or an account like it) could work without holding that word meanings are necessarily concepts (or, equivalently, that words encode concepts as a rule). In fact, by the end of this chapter, we will see that in some ways, other recent developments of Relevance Theory are not so far from the position I will outline here (Sperber and Wilson, 2015). If something like Relevance Theory is compatible with my suggestion that words do not stand in a reliable correspondence with concepts, then I think it really is a large benefit for that theory. If a theory of concepts or meanings implies that there is such a thing as JUSTICE, then that theory should be able to give an account of what properties JUSTICE has; JUSTICE has to be able to fit in with the rest of the theory's commitments; and JUSTICE should *explain* something. But that doesn't actually seem to be the case when we try and spell out properties of JUSTICE in terms of any given theory. So, if we can get by without positing JUSTICE, then we have solved quite a serious problem for theories of meanings and concepts in general. First, I give a brief overview of a traditional conceptual version of Relevance Theory (Sperber and Wilson, 1995; Wilson and Carston, 2007), and show how it is supposed to explain the interpretation of utterances by appealing to concepts. Then, I hope to show that Relevance Theory still works in spirit (better, in fact) if some of the burden of explanation is shifted away from concepts, and onto past usage episodes instead. Finally, I consider some objections and potential responses. In considering some of these objections, I also want to show that the suggestions I make in this chapter are not as radical as they might first appear: it is certainly possible to be a Relevance Theorist and dispense with JUSTICE.

8.2 Overview of Relevance Theory

Relevance Theory (Sperber and Wilson, 1995) is a general theory of human cognition with an emphasis on explaining utterance interpretation. It starts with the assumption that human cognition is specially adapted to maximise the *Relevance* of the results of processing stimuli. The Relevance of a stimulus is defined as a trade-off between the cost of processing it, and the benefits that processing it might bring. The benefits of processing a stimulus are called *Positive Cognitive Effects* (roughly; beneficial pieces of information; useful adjustments to a mental model of how the world is). So highly Relevant stimuli are those stimuli that produce lots of positive cognitive effects for relatively little processing effort. What exactly contributes to increased effort in processing a stimulus is vague and varied. But the kinds of thing Relevance Theorists have in mind regarding utterance interpretation are relatively uncontroversial assumptions such as: interpreting the sound [baʊ] as referring to a part of a ship as opposed to part of a tree will be easier in some contexts than in others, because of the nature of those contexts and the frequency with which each interpretation has been encountered before. Contexts are psychological constructs: they involve arrays of so-called *Contextual Assumptions*. Contextual Assumptions are pieces of information represented in memory (or perceptually available from the context of the utterance) that can be combined with information derived from a stimulus in order to make Relevant inferences. These Contextual Assumptions could, in principle, be almost anything that would help an interlocutor draw Relevant inferences. For example, in a context in which I respond to my supervisor's question of whether our 14.00 meeting is still going ahead, my utterance of "I just realised I left the oven on" might make the following Contextual Assumption 'accessible' to my supervisor:

Input: Lewis has said that he has left the oven on

CA1: Ovens that are left on can cause fires

This Contextual Assumption could, when combined with other such assumptions and the specifics of the conversation, yield the Relevant implication that I cannot attend the meeting because I intend to rush home and check that my house has not burned down:

CA2: People generally want to be sure that their house is not on fire

CA3: Attributing to Lewis the intention to check his house is not on fire instead of attend the meeting would provide an answer to my question as to whether the meeting is going ahead

Contextual Implications: Lewis cannot attend our 14.00 meeting because he will shortly rush home to check his house is not on fire; the meeting is not going ahead

These contextual implications are Relevant because they provide positive cognitive effects (they provide my supervisor with the information she wants) and because they were easy to arrive at given the context. On the Relevance Theory view, speakers exploit a bias in hearers to expect that their utterances will have an Optimally Relevant interpretation: interpreting the speaker's utterance with respect to certain constraints will give rise to a worthwhile number of positive cognitive effects for relatively little processing effort. Hearers assume that a speaker has designed their utterance such that the interpretation that is easiest and seems most useful (i.e. has sufficient positive cognitive effects) in a given context is precisely the interpretation that the speaker intended them to generate in that context.

The final element of Relevance Theory that concerns us here is what information an utterance itself actually contains (or makes accessible to the cognitive system). This information is cashed out in terms of the relation between words and concepts, and what properties concepts have. Sperber and Wilson's (1998, p. 189) description of what they take concepts to be is very much in line with my notion of unitary cognitive resources, outlined in Chapter 3: 'a concept… is an enduring elementary mental structure, which is capable of playing different discriminatory or inferential roles on different occasions in an individual's life'. They also endorse the view that thought is composed of concepts in the standard way: 'we assume that mental representations have a structure not wholly unlike that of a sentence, and combine elements from a mental repertoire not wholly unlike that of a lexicon'. However, Sperber and Wilson (1998) argue that the relationship between the number of words we have and the number of concepts we have is not one-to-one. Instead, we have a great many more concepts than we do words. When we have a concept but no corresponding word for it, it is an 'unlexicalised concept'. Crucially, Sperber and Wilson (1998, p. 43) do still seem to think that words encode concepts generally: 'a word which encodes a given concept can be used to convey… another concept… almost any word can be used in this way… it may so happen that the intended concept is the very one encoded by the word, which is therefore used in its strictly literal

sense'. So, on the traditional Relevance Theory approach to word meaning, words do encode concepts as a rule. This suggests that, for example, the word 'justice' encodes the concept, JUSTICE. The concepts encoded by words (and the context of the utterance) provide evidence that points towards a speaker's meaning, but this evidence doesn't fully determine a speaker's meaning all by itself.

An innovation of Relevance Theory is the idea that the concept encoded by a word is not ready-made to fit into any and all possible interpretations of any and all utterances that contain that word. Instead, the concept is a kind of file address for an array of logical and encyclopaedic information. This logical and encyclopaedic information is then used to construct an occasion-specific interpretation of the word; a so-called 'ad-hoc concept' (Wilson and Carston, 2007). To give an idea about how this works, consider a standard kind of example from the literature:

(1) I won't be having a drink tonight.

It is not hard to imagine a context in which an utterance of (1) is used to express the proposition that the speaker will not be having alcoholic drinks specifically, as opposed to making a blanket claim about not ingesting any fluid at all. Relevance Theory has a very neat explanation for this. The word 'drink' encodes the concept, DRINK. DRINK is a file address that points to, or makes available, logical and encyclopaedic information associated with DRINK. Encyclopaedic properties are mentally represented bits and pieces of information associated with (members of) the category that a concept stands for. Encyclopaedic properties are relatively unconstrained: the encyclopaedic properties that the encoded concept DRINK points to could include: that drinks come in various colours; that some drinks are alcoholic, an association between Irn Bru and the phrase, 'Scotland's other national drink', and so on. The unconstrained nature of these encyclopaedic properties is what gives Relevance Theory much of the desired flexibility in order to explain utterance interpretation. And, as an aside, I don't mean to suggest that this unconstrained-ness is a bad thing: it is clear to me that I know all of these things; that they all have something to do with drinks; that they might mediate my drink-related behaviour in some contexts, and so it seems plausible to me that they would make up part of the resources associated with my DRINK concept. Logical properties, on the other hand, are representations of information that provide the basis for deductive reasoning about category membership and inferential relations between concepts ('meaning postulates'). One such logical property pointed to by the concept DRINK might be an 'elimination rule' that takes the concept DRINK as input, and returns a property such

as EDIBLE FLUID. Logical properties license deductive and other inferential reasoning processes, but they are not supposed to provide necessary and sufficient conditions for category membership, or definitions of extensions (Sperber and Wilson, 1995, p. 92). The important difference between encyclopaedic and logical properties is that according to this picture, there is nothing about experience of Irn Bru that is *constitutive* of the DRINK concept. It could turn out that in the near future, scientists decide that some chemical contained in Irn Bru is deadly poisonous. However, gaining an appreciation of this fact would not change DRINK fundamentally, because its logical properties are still in place. I assume that it is in this way that the DRINK concept is able to play 'different discriminatory or inferential roles on different occasions', while still being characterisable as an 'enduring' mental structure.

Relevance Theory holds that, quite generally, encyclopaedic and logical properties of the concepts encoded by content words are manipulated by the hearer according to constraints of Relevance in order to produce ad-hoc, occasion-specific concepts. In the case of (1), uttered just as a particularly wild party is about to commence, certain Contextual Assumptions would be accessible, such as "people stereotypically drink alcohol at wild parties". The hearer would then 'narrow' the encoded concept DRINK by focusing on certain of the encyclopaedic properties that it points to (those encyclopaedic properties that describe alcoholic drinks). This results in the ad-hoc concept, DRINK*. DRINK* denotes not all drinks, but only those drinks that are alcoholic. This decision on the part of the hearer is licenced by the constraints of Relevance discussed above and the context of the utterance. The context of the utterance has the same kinds of effects as in the oven case above. Suppose (1) was uttered in response to Smith, who is proffering a can of Oranjeboom: narrowing DRINK to DRINK* provides Smith with an interpretation that is Relevant in this context: the speaker does not want a can of Oranjeboom because the speaker will not be having any DRINK*s, and DRINK* includes Oranjeboom in its denotation. Note that on this account, if a speaker of (1) downs a glass of water immediately after making their utterance (in the party context), then on the Relevance Theory account the proposition they have expressed is still true. That is because water is not denoted by DRINK*; only alcoholic drinks are. If (1) was uttered in a different context, say one in which body builders are preparing for a competition by intentionally dehydrating themselves, then 'drink' could be used to express a different ad-hoc concept (perhaps an ad-hoc concept closer to the one encoded by 'drink'). In that case, the speaker would express a different proposition even though the sentence type they have uttered is the same in both scenarios.

So, that is roughly how concepts figure in utterance interpretation on the Relevance Theory account. To summarise: words encode concepts. Concepts point to, or make accessible, encyclopaedic and logical properties. A subset of these encyclopaedic and logical properties is selected on the fly in utterance interpretation in order to generate ad-hoc concepts that promote Relevant inferences.

8.3 The meaning of 'dog' and 'justice' in Relevance Theory

Let us now consider two utterances from the Relevance Theory perspective, one of which turns on the concepts encoded and conveyed by a use of the word 'dog', the other turning on those concepts encoded and conveyed by a use of the word 'justice'. A familiar pattern emerges: in the case of 'dog', the relation between the interpretations, the theoretical constructs, and their properties is straightforward enough. However, in the case of 'justice', I argue that things are much less clear. Suppose that (2) is uttered by someone (A) wanting to give advice to a friend (B), who is scared of their new neighbourhood. Relevance Theory has a nice account of why we have certain intuitions about what A is likely to mean, and what proposition they are committing to.

(2) A: If you've got a dog then you don't have to worry about being burgled

We don't take A to mean that just any old dog would do. If B comes into possession of a 2-month-old mute, blind Chihuahua, gets burgled, and then complains to A, then we might feel that it is B's behaviour that is strange, as opposed to A's. Ad-hoc concepts can explain why this is. Setting the rest of the utterance aside and focussing just on the word 'dog' and its encoded concept, we could say that that the intended interpretation of (2) involves a narrowing of DOG so that encyclopaedic properties describing loyalty, size, fierceness, intelligence, and so on, are emphasised. The result is DOG*: an ad-hoc concept that denotes those dogs that are good at guarding houses, and does not denote 2-month-old mute, blind Chihuahuas. A Contextual Implication that B could draw in this case could be something glossable in English as, "A is suggesting that I should buy a DOG*".

Now, consider an utterance of (3), uttered by a campaigner (C) who is trying to combat the recent fashion for subjecting strangers to insults over the internet:

(3)

D: Why has Charity Z made so little progress in making the internet a safer place?

C: It's hard to get justice for online abuse.[10]

My own intuition about this example is that the *proposition* that C is committing to is not as determinate as in the case of (1) or (2). And, although this is just an intuition, you might wonder why that should be, if reconstructing the speaker's meaning in (3) is a matter of building a proposition from encoded concepts and/or ad-hoc concepts. If concepts function as categories and compose into thoughts, and words encode these concepts, then it seems to me that we should expect the proposition that we attribute to a speaker to be relatively clear.[11] But, as a rough approximation, suppose that C is trying to express a proposition that might be glossed as: "it is difficult to get society to punish people who engage in online abuse". Even if you think this gloss characterises *the* proposition that is definitely the one that C intends to communicate, I still don't think it's straightforward to gloss what the *specific* contribution of the word 'justice' is to the interpretation of this utterance. A suite of problems arises here. Sperber and Wilson (1998) suggest that a 'strictly literal' use of a word occurs when it is used to convey the very concept that it encodes, as opposed to an ad-hoc concept. So we should also be able to say what would count as a literal use of the word 'justice'. Is (3) a literal use of the word 'justice'? There does not seem to be a principled way of answering this question. In some contexts, the word 'justice' can be used to refer to or describe situations that involve capital punishment, but our campaigner is not necessarily indicating that she believes that Twitter arguments could or should be resolved in this way. So, as with 'dog' in (2), the speaker here does not intend to denote just *anything* that could be felicitously described with the word 'justice'. She has a specific kind of thing in mind; perhaps a hefty fine for perpetrators, newspaper articles that draw attention to the plight of victims, and so forth. But it doesn't seem like this "kind" of justice has any more claim to literality than capital punishment.

A potential objection to this interpretation of C's utterance is that it assumes that C is using the word 'justice' (and, therefore, the concept JUSTICE) to talk about things which more plausibly fall under other concepts. For example, a hefty fine seems like a prime example of a punishment (and so is plausibly denoted by the

[10] Some commentators on drafts of this chapter have complained that this particular sentence sounds unnatural. However, it is a slightly modified version of an attested instance of the word 'justice' taken at random from a newspaper article search ("Why it's so hard for women to get justice for online abuse"). That being the case, I think we should expect a theory of language interpretation to be able to account for it. In any case, I do not think too much turns on the specific example we will consider here.

[11] I return to this issue below.

concept PUNISHMENT), but a punishment is not necessarily the same thing as justice. My response is that I completely agree: here we seem to have a case where the word 'justice' is used in such a way as to token mental resources that don't seem to correspond neatly to the alleged concept, JUSTICE. In fact, this is another example of the phenomenon I discussed in sections 3.5 and 3.6: any time we try to use the concept JUSTICE to explain human behaviour or cognition, some *other* theoretical entity ends up doing all of the explanatory work. Still, you might further object that this *can't be* a felicitous interpretation of the utterance, because it doesn't contain the concept, JUSTICE. My response here is that this simply begs the question: the issue at stake is precisely whether interpretations of utterances containing the word 'justice' do in fact require the concept, JUSTICE. My point is that these interpretations do not require us to posit the concept, JUSTICE, and so I won't grant that assumption.

Still, you might object, why can't we say that the concept JUSTICE makes available some notion of 'appropriate punishment' via its encyclopaedic properties? Then, surely, there *is* a role for JUSTICE to play in utterance interpretation. I want to spend a while considering this idea, because I think that it raises issues that are deeply troubling for traditional conceptual relevance theory, and that the past-usage account I will sketch out later may provide a way around such difficulties. Let's suppose that here we have another situation in which we need to derive an ad-hoc concept, JUSTICE*, from an encoded concept, JUSTICE. It seems to me that we immediately run into the same issue facing Barsalou, Fodor, et al., as discussed in chapter 3, but in a slightly different guise. We now have to provide an account of the concept encoded by the word 'justice', and so there should be some plausible encyclopaedic and logical properties of JUSTICE that help to explain how we interpret the utterance in the way that we do. We do not, *as a matter of course*, take the campaigner to be indicating that online abusers should receive capital punishment, and instead we reliably come to a different interpretation. The Relevance Theoretic notions of ad-hoc concepts, encyclopaedic properties, and logical properties seem designed to account for situations like this.

However, for me, it is immediately *much* more difficult to come up with straightforward encyclopaedic properties for JUSTICE than for DRINK or DOG. Perhaps some candidates for encyclopaedic properties made available by JUSTICE could be: that historically, people were hung for committing certain crimes; that some people have been falsely imprisoned and subsequently had their sentences overturned; that if you are wronged then cultural norms dictate that you are entitled to some redress, and so on. Perhaps some such properties are focussed on in order to generate the ad-hoc concept, JUSTICE*, which denotes, say, hefty fines and

advocacy in periodicals. The problem here is that these are, arguably, very different kinds of properties than the ones pointed to by DRINK and DOG. The encyclopaedic properties pointed to by DRINK and DOG were properties of drinks and dogs. But it does not seem to me that the knowledge that historically, people were hung for committing certain crimes is a 'property of' justice. Likewise, the view that the wronged should be able to seek redress is not a 'property of' justice. Furthermore, these 'properties' seem like they contain very complex propositional attitudes ("*I believe that* if someone is wronged then they should be able to seek redress"). A propositional attitude is a paradigm case of a thought, and in the general model, thoughts are built out of concepts. This is the point that I think might cause some major issues for traditional conceptual Relevance Theory. If we allow that encyclopaedic properties made accessible by concepts can themselves be built out of concepts, then we have the danger of an infinite regress, and no obvious way of breaking it.

Here is how we get trapped in an infinite regress. Recall that encyclopaedic and logical properties are supposed to license Relevant implications and inferences. Ultimately, the point of all of this is that we want to explain how hearers generate interpretations of utterances. The hearer needs to work out what it is that a speaker means. Here, the speaker has said:

(3) It's hard to get justice for online abuse.

Focussing just on the word 'justice' for the time being, traditional Relevance Theory has it that the word 'justice' activates the concept, JUSTICE. The concept, JUSTICE, is a mental file address that makes an array of encyclopaedic and logical properties available. Let's suppose that one such property is a propositional mental representation, which we can gloss as:

(4) (JUSTICE IS) SOCIETY'S IMPLEMENTATION OF FAIR TREATMENT OF ITS CITIZENS.

Well, if this mental representation is *itself* a propositional structure built out of concepts (SOCIETY, IMPLEMENTATION, FAIR…), then so far we haven't gotten any closer to working out what the speaker meant by their use of the word 'justice'. The hearer needs to work out *how* they should modulate JUSTICE, to generate the ad-hoc concept, JUSTICE*. But concepts themselves are just file addresses, on the traditional Relevance Theory account. In order to cash out (4), a hearer needs to consult some encyclopaedic properties made available by the concepts in (4). One

way of putting this is that it's reasonable to ask: what counts as 'fair' in this case (what are the encyclopaedic properties activated by FAIR)? But if these *subsequent* encyclopaedic properties are again just propositional structures built out of more concepts, then all a hearer has achieved is the activation of *another* set of file addresses that do not themselves provide the resources that the hearer needs in order to work out what the speaker of (3) meant by their use of the word 'justice'. If we simply define concepts in terms of other concepts, as seems to be the case in (4), then it is unclear how any of this hypothetical cognitive processing eventuates in "meaning" (or, "interpretation"). It seems like all we get is file names pointing to more file names, without ever actually getting to the "stuff" that the files contain. The problems do not stop here. Remember that I am only using the alleged concept JUSTICE as an example. I think that at least some other alleged abstract concepts might also be vulnerable to the criticisms I level against JUSTICE (as we saw with DEMOCRACY at the end of Chapter 3). The elements of (4) are supposed to be concepts themselves, so we can ask: are these concepts any less problematic than JUSTICE? For example, in order for the activation of the concept IMPLEMENTATION to provide the hearer with the resources they require to interpret the speaker, IMPLEMENTATION also needs to activate some Relevant encyclopaedic properties. It just isn't clear to me that we're going to be able to say much about what these properties are.

As I just pointed out, if these other encyclopaedic properties are built from more concepts, then we are getting further and further away from what we initially wanted, which was an explanation of how the ad-hoc concept JUSTICE* is derived from the encoded concept, JUSTICE. And this leads me back to one of my original arguments against the explanatory value of JUSTICE as a theoretical posit (presented in Chapter 3). Suppose that we grant the conceptual Relevance Theorist that (4) characterises a legitimate encyclopaedic property of JUSTICE, and furthermore, we even grant that the concepts in (4) somehow bottom out in such a way that the interpretation of utterance (3) is explained. Well, even if *that* were true, once again we have a situation in which JUSTICE itself isn't doing any of the explaining! The concept JUSTICE does not appear in (4). Instead, it's *other* concepts that explain how a hearer ultimately interprets a speaker of (3), namely SOCIETY, FAIR, TREATMENT, and so on. All the concept JUSTICE seems to do is activate other concepts. Why should we bother stipulating this? It seems to me that we could easily allow that the word 'justice' activates the Relevant concepts immediately, without an intermediate step in which a seemingly content-less concept is activated that does not actually aid interpretation.

Before we move on, I want to address a potential further worry that one might have in light of the arguments I just presented, but which I think is unfounded. If the encyclopaedic properties made available by concepts can themselves be made of concepts, then we might worry that conceptual Relevance Theory wouldn't be able to explain the interpretation of *any* utterance. Why shouldn't we also worry that the concept DOG, activated by the word 'dog', suffers from the same regress problem as JUSTICE? The reason I think this worry is unfounded is that Relevance Theorists allow that there are some encyclopaedic properties that don't seem to be propositional structures built out of concepts:

> [encyclopaedic information can include] conceptually represented assumptions and beliefs… **and also**… imagistic and/or sensory-perceptual representations… kinds of bodily movements… idiosyncratic information (episodic memories) based on one's own observations and experiences…
>
> (Carston, 2010, p. 246, emphasis mine)

So, when it's *not* the case that an encyclopaedic property just is a propositional structure built out of more concepts, then there could be some mental representations that can licence an inference about what a speaker might mean. As we saw in the previous sections in this chapter, it seems much easier to come up with plausible non-propositional encyclopaedic properties for DOG than JUSTICE, and so for that reason the traditional conceptual Relevance Theorist hopefully does not have to worry too much about utterances containing the word 'dog'.

I take the above to show that conceptualist Relevance Theory has some explaining to do when it comes to what properties JUSTICE points to. However, I want to stress that I don't think that, on its own, this forces the conclusion that conceptualist Relevance Theory is wrong about words and concepts. The conceptual Relevance Theorist has at least two possible ways of dealing with the issues I have outlined here. Firstly, the conceptualist Relevance Theorist could take up the gauntlet and provide convincing derivations of interpretations of utterances involving the word 'justice' with reference to plausible encyclopaedic and logical properties. I hope to have convinced you that this will be difficult, but I certainly haven't proven that it is impossible. Secondly, the Relevance Theorist could simply accept that some words encode concepts that point to encyclopaedic and logical properties that are ineffable: we can't gloss what kinds of encyclopaedic properties JUSTICE points to because they're not the kinds of things that it's possible to render in English. I am not sure if

that's compatible with all of the other commitments that one might want to hold as a Relevance Theorist, and I suspect that, in any case, most would find that move unpalatable, but the option is there.

8.4 A sketch of a non-conceptual account of word meaning

Now, I will attempt to sketch out how an alternative view might explain how we interpret (2) and (3) without holding that the only kind of information a word makes available is the concept it encodes. To reiterate, I think the right question to ask when it comes to explaining the phenomenon of communication (which is a major goal of Relevance Theory) is not, "what is the meaning of the word X?", but instead, it is "what cognitive resources could we attribute to Smith in order to explain his understanding of a use of the word X?". Instead of content words just encoding concepts, I think it is plausible that, over time, starting early in ontogeny and continuing throughout a lifetime, humans build up a library of episodes of the use of a lexeme or phrase. Again, the goal here is not to convince you wholeheartedly that this account is correct, but merely to show that there are viable alternatives to the conceptual view of word meaning. Reconstructing the proposition that a speaker intends to express is a matter of building an interpretation on the basis of information contained in these stored past usage-episodes, as well as information present in the discourse context.

A pressing concern here is: what is the nature of the 'information' stored in these past-usage episodes? Whatever it turns out to be, it had better not just exclusively be 'concepts', because then my account will collapse into the conceptual word meaning account. In this first stab at outlining an alternative, I see no reason to put all that many limits on what information could be stored as a result of a past-usage episode[12], or, if you prefer, on what ways this information could act as a constraint on interpretation. A large part of it could simply be a record of those inferences and responses that an interlocutor seemed to accept on previous occasions of use, instantiated as patterns of brain activity. Other kinds of information could consist of perceptual representations of what happened to be in the visual field on previous occasions of use, partial imprints of affective states on previous occasions of use, and so on. Couching things in Relevance Theoretic terms, past usage information could be just any subset of the state of the mind-brain that was involved in producing Relevant results when words and phrases were heard or used. It strikes me that this

[12] From now on, I will use the phrases 'past-usage episode' and 'stored past-usage' as a shorthand for 'information that is stored as a result of a past-usage episode'.

kind of information is no more or less constrained than what Relevance Theorists seem to have in mind when they talk of the encyclopaedic properties that are made available by concepts, although they are of a different nature. Concepts are supposed to function as cognitive stand-ins for categories, and do double duty as the kinds of things that compose into thoughts, but I am not suggesting that past-usage information necessarily determines the extension of a category or is necessarily available to the procedures that produce propositional thought. And so there are good reasons to hold that a set of past usages does not constitute a concept. A concept is a unitary cognitive resource that composes into thoughts in a systematic way, and it must also support cognising about and categorising entities as such. A stored set of past usages of the word 'justice' does not seem to correspond to a concept on anyone's account, because there is no category that it picks out. On this kind of view, a stored past usage really could just be 'some sub-state the mind-brain was in when it encountered a word'. Relevance Theory hypothesises that humans are geared towards maximising the Relevance of processing stimuli, and so over time, those subsets of the mind-brain that did actually produce Relevant results will stabilise to a degree, and noise will tend to average out.

So, very broadly, anything that could be represented as a memory trace could in principle be a component of a stored past-usage of a word. Over time, Smith could amass a large and varied pool of information of this sort captured from encounters of usages of the words 'dog', 'justice', and so on. In a given context, some pieces of this past-usage information will be more or less accessible as a function of factors such as how structurally similar an incoming linguistic string is to previously encountered linguistic strings, how easy they are to combine with contextual assumptions in order to derive inferences, and all of the factors that Relevance Theory, or any other theory for that matter, incorporates. I think that Relevance Theoretic principles could operate on these past usage episodes in more or less the same way as they supposedly operate on concepts. Speakers are still assumed to be trying to express propositions in an Optimally Relevant way, Contextual Assumptions are still more or less accessible as a function of discourse context, and Contextual Implications still follow from combining the information extracted from processing an utterance with Contextual Assumptions.

8.5 The meaning of 'dog' and 'justice' in a non-conceptual account of word meaning

So, on to utterances (2) and (3). Here, I just want to show that the past-usage account does a reasonable job of reconstructing how a hearer generates an interpretation of these utterances. If you accept that the past-usage account *does* actually do a reasonable job, then objection 2 is dealt with. Consider (2) again:

(2) If you've got a dog then you don't have to worry about being burgled.

In order to be just as viable as standard Relevance Theory, the past-usage account has to be able to explain what the role of the word 'dog' is in the interpretation of this utterance in terms of information stored in past-usages, instead of an encoded concept. I think that "in the real world", the kinds of chains of cause and effect, inference, input, output, contextual influence, and so on could potentially be both very complex and not necessarily amenable to glossing in English. So I just want to show the *kind of thing* that could take place on the past usage account, but I don't take myself to be providing 'the' full explanation of how any hearer might interpret (2). With that caveat, B's interpretation might be reconstructed in the following way:

Contextual assumption: B is worried about being burgled in her new neighbourhood (NB: this was given)
Input: A's utterance contains, among others, the words 'dog' and 'burgled' arranged syntactically in a certain way

B has stored past-usages of the word 'dog', encountered in such contexts as seeing a 'Beware of the dog' sign, among many others. Some of these past-usage encounters occurred while a particularly fierce-looking animal was making a lot of noise.

Contextual assumption: B wouldn't like to be in the shoes of a burglar confronting *that kind* of animal (made accessible by a combination of the context and past-usage episode information)
Contextual assumption: People acquire such animals (and signs) to put potential burglars off
Contextual assumption: People who have put potential burglars off are less worried about being burgled

Contextual implication: if B bought the kind of animal she associates with 'Beware of the dog' signs, she could worry less about being burgled.

Contextual implication: A is suggesting B buy *that kind* of animal

In this way, B can come to an interpretation that agrees with our intuitions about what A meant. But explaining this interpretation didn't require the assumption that the cognitive resources made available by the word 'dog' begin and end with the concept, DOG. Note that this does not imply that A or B *lack* a DOG concept. It's just that A and B can understand each other verbally in a way that isn't *necessarily* mediated by something that we would want to call a concept. Now is a good time to stress this important feature of the account I am sketching here. I am not claiming that concepts *never* feature in interpretations of utterances. Rather, I am claiming that concepts are not a *necessary* feature of an interpretation of an utterance. Given certain contexts and utterances, it could be that a successful interpretation does involve a mental structure built out of concepts. So, it could *contingently* be the case that a successful interpretation of a particular utterance does involve something we would want to call a concept. The difference, to the extent that there is one, between the words 'dog' and 'justice' is just that in the case of 'dog', a successful interpretation might typically involve a tokening of a hearer's DOG concept (along with information stored in past usage episodes of the word 'dog'). However, in the case of 'justice', any concept that does get tokened will not correspond to a unitary cognitive resource, JUSTICE. And, in any case, for interpretations of utterances containing either 'dog' or 'justice', past-usage information plays a large role in generating interpretations. I accept that right now, you might be concerned about exactly *how* this past-usage information affects utterance interpretation, and interacts with contextual assumptions. I will address these issues below I consider possible objections to the position I am sketching here.

For now though, on to utterance (3):

(3)

D: Why has Charity Z made so little progress in making the internet a safer place?

C: It's hard to get justice for online abuse.

Assuming that C's interlocutor (D) has had a typical exposure to the word 'justice' throughout her life, then there are any number of previously drawn inferences, memory traces, and occasion specific Contextual Assumptions that she could draw on in order to generate an interpretation of C's utterance that is Relevant to her:

Input: C's utterance contains, among others, the words 'abuse', 'get', and 'justice' arranged syntactically in a certain way

> D has stored past-usages of the word 'justice' that include: it being used to label a state of equity; D has stored past-usages of the *phrase* 'get justice' that include it being used to describe situations in which an attempt was made to right a wrong; D has stored past-usages of the word 'abuse'…

> Contextual assumption (motived by past-use): Situations labelled by the word 'abuse' are iniquitous
> Contextual assumption: Iniquity should be remedied
> Contextual assumption: (motivated by past-use): to 'get justice' may involve remedying iniquity
> Contextual assumption: People have different ideas about what response is appropriate when remedying iniquity
> Contextual assumption: People who work for Charity Z tend to have a certain kind of view about remedying iniquity
> Contextual assumption: People who work for Charity Z don't generally believe in capital punishment
> Contextual Implication: Remedying iniquity in this case does not involve capital punishment, but some other kind of societal response
> Contextual Implication: C believes that it's hard to effect *that kind* of societal response regarding online abuse
> Contextual Implication: If it were the case that something is hard, that would be a reason why it hasn't been achieved…

Note that the contextual assumptions suggested above are just as available to the conceptual Relevance Theorist (modulo mentions of past-usage) as they are to me. But this does not get the conceptual Relevance Theorist off the hook, because these contextual assumptions and implications were supposed to licence the generation of an ad-hoc concept, JUSTICE*, and it's at *that* point that I argue the account runs into trouble. On the past-usage account, the contextual assumptions and resulting implications serve as a gloss of what D's interpretation of C's utterance is likely to be. These contextual assumptions and implications are Relevant for the same reasons that anything is Relevant according to Relevance Theory. They were easy to derive given the occasion of use, and the information made accessible by the linguistic string

C uttered. They also provide D with the cognitive effects she is after: a certain kind of societal response being hard to effect would provide a reason why Charity Z has not made the progress she desires. Again, although the details are obviously somewhat sketchy, the resulting interpretation seems felicitous enough given the context and what C uttered, and there is a clear path from the input D receives, to the interpretation she generates, via general pragmatic principles. We could tell the same kind of story for the other words in the utterance ('hard', 'abuse') and the past usages of those words are likely to contain information that will be mutually reinforcing with regards to the interpretation that D generates. The past-usage account can explain how we generate interpretations of utterances without holding that we must interpret the linguistic content of utterances only by composing concepts encoded by words into propositions. Now I shall consider some objections and potential responses.

8.6 Objections to a non-conceptual account of word meaning

Perhaps the most pressing objection stems from a worry about just how much work these Contextual Assumptions are doing: where do they come from, and what are they made of, such that they can promote the kinds of inferences that produce the interpretation that B and D come to? Are they, for instance, thoughts made of concepts? If they are not thoughts made of concepts, then how do they figure in inferences at all? I have a few responses to these lines of inquiry. The first thing I want to stress is that I am *not* claiming that there are no such things as concepts, or indeed that there are no such things as thoughts built from concepts. I accept the explanatory power of concepts as a psychological construct, and I think that in many cases, words of English are likely to pick out concepts (such as with DOG). I also accept that the only successful model of the mind that anyone has come up with so far assumes that thoughts have parts. In this thesis, my point is just that *JUSTICE isn't one of these parts.* So it could well be that, in the course of interpreting a specific utterance, a hearer does entertain contextual assumptions which licence Relevant inferences, and these contextual assumptions are made of concepts. The issue I have been trying to raise in this chapter is just that, if it turns out that a Relevance Theory derivation requires the concept JUSTICE, hidden away in a contextual assumption or elsewhere, then we run into the (serious) problems I have been expounding. One major benefit of the past-usage account is that using the word 'justice' in a derivation doesn't require a commitment to the existence of the concept, JUSTICE. It simply requires a commitment to there being some memory traces of past-usages of the

word ‘justice’ that could plausibly motivate some Relevant cognitive effects. We gloss in English the effects of these memory traces, and the result of the processes they play a part in, because we have to gloss them in *something.* But the gloss isn’t supposed to mirror exactly the semantical structure of these cognitive processes and representations.

That being so, a Relevance Theorist might still worry about the questions raised in the previous paragraph. The traditional story has it that words encode concepts, concepts stand for categories, and concepts compose into thoughts. This goes a long way towards explaining how interlocutors draw inferences in conversation (the speaker said they want an X; a Y is a kind of X; so the speaker might be satisfied if I give them a Y…). If at least some of what we call an ‘interpretation’ of an utterance is *non*-conceptual past-usage information, then how does this past-usage information feature in inferences, or affect the results of inferential processes? Furthermore, how do we work out what a speaker has asserted, unless this assertion is representable as a propositional mental structure built from concepts? If you are worried about these questions, then my response may make you feel short-changed. As I mentioned in section 8.1, at least one modern development of Relevance Theory makes *somewhat* similar moves to the ones I have been trying to motivate in this chapter (Sperber and Wilson, 2015). Here is a choice quote:

> Not all inferences involve step by logical step derivations of explicit conclusions from explicit premises… What [sometimes] happens… might be better described as changes in patterns of activation, none of which would properly speaking amount to the fixation of a distinct credal representation, but the totality of which would correspond to the formation of an impression… More generally, many (if not all) inferences can be described not as more or less standard logical derivations but as competitions between alternative conclusions (it will rain/it won’t rain, let’s go for a walk/let’s not, and so on). The winner of such competitions is determined by activation or inhibition caused by brain states that represent information in all kind of ways.
>
> (Sperber and Wilson, 2015, p. 137)

Here, Sperber and Wilson appeal to conscious and unconscious collections of results from the processing of a stimulus, which we cannot paraphrase in any metalanguage.

Furthermore, they allow that such mental representations can play a causal role in utterance interpretation. Sperber and Wilson also argue in this paper against the idea that hearers and speakers *necessarily* entertain distinct, unitary propositions, and they explicitly state there are cases in which 'pinpointing a proposition that would constitute the speaker's meaning is difficult or impossible' (Sperber and Wilson, 2015, p. 146). Now: as far as I can tell, Sperber and Wilson would not endorse my past-usage account, and it seems to me that I take a more radical position than they do regarding the (in)determinacy of propositions. However, they clearly allow that non-conceptual sources can have effects on utterance interpretation and inference. So if you're a Relevance Theorist, as Sperber and Wilson certainly are, the idea that there *are* such cognitive entities and effects can't on its own be grounds for claiming that my position is incompatible with Relevance Theory. And again: I am not trying to argue *against* Relevance Theory. I am trying to find a way of reconciling Relevance Theory with the issues that the alleged concept JUSTICE causes, because it is a popular and (at least partially) successful theory of communication. My solution is to avoid positing JUSTICE. I have tried to show in these sections that if we avoid positing JUSTICE, Relevance Theory can work just as well, while getting shot of some serious problems. And furthermore, where concepts *do* enter the picture on my account, the machinery of ad-hoc concept generation is still as powerful as it ever was.

Another potential objection is that you might be worried by my reliance on the phrase 'arranged syntactically in a certain way'. You might wonder what exactly the role of this syntactic arrangement is. I take my proposal to be neutral with respect to any particular syntactic theory. In the account I am sketching, the syntactic structure of a linguistic string acts as a constraint on what a speaker could have meant by uttering something in much the same way as in other theories of utterance interpretation. I have in mind here such constraints as: in English, there is a certain structural element called a subject, and subjects are more likely to have certain message-relevant properties than others. These kinds of constraints could make some past-usage information for a particular word or phrase more Relevant than other such information, but the existence of past-usage information itself doesn't seem to me to be incompatible with whatever syntactic structure a theorist wants to assign to a linguistic string. In fact, there is at least one recent study of the interface of syntax and word meaning that seems to fully endorse a past-usage style position and explicitly denies the traditional conceptual position (Wechsler, 2015).

A further potential objection related to the issues just discussed is, if we endorse the past-usage account, then we lose an important link between the structure and semantics of natural language sentences and the structure and semantics of

thoughts. Conceptual Relevance Theory comes with a very compelling story about the relationships between the structure of linguistic strings and the structure of the thoughts that an interpretation of them consists of. Interpretations of utterances in Relevance Theory are made up of encoded and ad-hoc concepts that compose to form propositions. So the thought that a hearer generates in response to processing an utterance has more or less the same structure and content (or, the same *kind* of structure and content) as that which they assign to an incoming linguistic string. This is considered to be a nice benefit of all theories that hold that words encode concepts: both thought and natural languages are held to be productive, systematic and compositional, and so it's generally held that the structure of a linguistic string and the structure of the thought that serves as its interpretation mirror each other in an interesting way. If you hold that words encode concepts, then you have an explanation of how these structures do mirror each other: word meanings and parts of thoughts compose in the same way because they're the same things. However, on my account, it's less clear how the 'meaning' of an utterance is composed from the meanings of its parts. This is because on my view, individual words don't contribute concepts that slot into a syntactic structure; instead they contribute information derived from past usages in a way that is constrained by syntactic structure.

However, I think this language-thought-structure objection rests on an assumption that we should be much more cautious about accepting. The assumption is that natural languages (although almost invariably the discussion is just about English) *have* a compositional semantics independently of any particular semantic theory *of* a language. It's taken to be axiomatic that 'English is compositional'; we just have to work out in what way it is in fact compositional. But it seems to me that this gets things backwards. Semanticists have set themselves the task of constructing a formal system that will take any grammatical (English) sentence as input, and produce a truth condition as an output. This truth condition should correspond to our *psychological* intuitions about what the right interpretation of the sentence is. One way to get started on this interesting task is to assume that the 'correct' (read: congruent with intuition) truth condition of a sentence is *computable* from the words contained in that sentence, along with a syntactic theory that specifies structural relations between those words. Put another way, semanticists are constructing a very interesting *model.* Compositionality is a *modelling assumption* that makes the problem of coming up with a formal system that can produce the right truth conditions from given linguistic strings tractable. But it is a completely separate thesis that this modelling assumption (or any of the other modelling assumptions the semanticist wants to make) has any psychological 'reality' with respect to what properties

'language' as a psychological object turns out to have. All that Fodor's language of thought hypothesis requires is that it's *thought* that is compositional, and likewise, Barsalou claims that simulators produce representations that can enter into structural relations with each other, but neither of these theories is supposed to be 'about English'. And I think this is more or less the conclusion that Fodor (2001) himself came to when he talked of 'the more or less patent uncompositionality of English', and it's being the case that 'thought, rather than language, has content in the first instance'. So the thoughts that interpreting a linguistic string *provokes* could be compositional, but that does not require that an utterance *itself* is compositional, whatever that might mean. And nor does it require that the *bits* that compose in thought have direct analogues in the surface structure of linguistic strings. A theory of concepts and/or thought should not be a theory of English semantics, unless that's an explicit commitment on the part of the theorist. I am suggesting that we avoid making this commitment unless some evidence suggests that we have to.

There are two more objections that might look dangerous, but I think the past-usage account can deal with them. The first objection is that the past-usage account is profligate: perhaps it seems very *im*plausible to suppose that the human mind-brain could store a library of episodes of past usages for every content word in a vocabulary containing tens of thousands of items. This line of attack can be deflected by the observation that the very same 'criticism' applies to any conceptualist's account. There is no reason to suppose that encyclopaedic and logical properties associated with every concept encoded in an adult human vocabulary would be more space efficient than a library of past usage episodes. The last objection I will consider is that: if we understand people by way of past usages of words, then that seems to imply that the proposition that a speaker intends to express and the proposition that a hearer reconstructs might be different. They might be different because the speaker and the hearer will have different libraries of past usages for any given word, and so they must be generating interpretations by using slightly different sets of tools. So *that* implies that we could never 'truly' grasp the propositions that our interlocutors are trying to express. *Concepts* might still be public, but we don't have a way of making sure that libraries of past usages are public. Doesn't that seem somewhat incongruent with our day-to-day experience, in which we seem to understand each other very well? I have two potential responses to this objection. The first potential response is to simply accept this result as a consequence of the view (and, actually, this is the response I prefer). It could be that, as a matter of fact, speakers and hearers will always entertain at least slightly different propositions. However, human beings generally communicate in order to achieve goals, and so they will design their

utterances in order to achieve those goals, and if a hearer gives a response that seems like it is more or less what the speaker was hoping for or expecting, then a speaker can be satisfied that at the very least, their hearer *appears* to have understood them. If we want to explain communication, then do we need to suppose more than this? It could be that you are still deeply sceptical of this result. In that case, my second potential response to the objection is that, assuming that speakers and hearers really *do* grasp the same proposition (whatever that might amount to), perhaps this could be explained by regularities in human brains, biology, and the environment. I have in mind here such regularities as the fact that the table-shaped collection of particles sitting in the corner of my living room looks similar enough to most (English-speaking) humans that they have all decided to use the same word to refer to it.

The past-usage account also has the benefit of being empirically falsifiable, in theory. For example, one or both of two results would be fatal for my account. If it could be shown that stored episodes of usages of a word could not, in principle, explain how we generate our interpretations of what other people mean when they use that word, then my account will be false. And if it could be shown that the psycholinguistic evidence is incompatible with there *being* stored episodes of usages of the words, then my account will be false.

8.7 Summary of Chapter 8

I will now clean up some loose ends. One major result from the preceding discussion is that in terms of interpreting an utterance, the word ‘dog’ is not much different to the word ‘justice’: interpretation in both cases occurs largely on the basis of non-conceptual information instantiated in past usage episodes of those words. I think that is a plausible story for all content words. The only connection (albeit an important one) between the word ‘dog’ and the concept, DOG, is that the word ‘dog’ may often be used by a speaker in such a way that it is likely that a hearer’s DOG concept is tokened. But tokening a DOG concept is just one way in which the word ‘dog’ might be used, and this is not a necessary component of all interpretations of the word ‘dog’. As I have been trying to argue, there is *no* connection between the word ‘justice’ and the concept, JUSTICE, because there is no such concept. Interpreting an utterance that contains the word ‘justice’ may contingently involve unlexicalised concepts that do not correspond directly with any lexeme of English. But in both cases (‘justice’ and ‘dog’), non-conceptual information stored in past usage episodes of these words plays a large role in generating interpretations of utterances which contain these words. One last worry with the theory I have sketched

in this chapter concerns what concepts are involved in thoughts that we try and express by using the word 'justice'. For example, you might think a thought expressible with an utterance of the sentence, "justice has been done". You might then suppose that this thought must have a semantics and structure that looks something like this: JUSTICE HAS BEEN DONE. I am arguing that this supposition is incorrect: the structure and semantics of thought do not necessarily map on to the structure and semantics that a compositional semantic theory assigns to English sentences. When you try and express a thought with the utterance, "justice has been done", you are simply using these words because the past usage information associated with them is such that you think these words are most likely to get a hearer to grasp the thought you are trying to express. However, this does not require that your conceptual repertoire contains the unitary cognitive resource, JUSTICE.

To summarise: in this chapter, I offered my response to objection 2 to my claim that not all words of English pick out concepts. In Chapter 3, I argued that there is no such thing as the concept, JUSTICE, because JUSTICE does not help us explain any human behaviour or cognitive processes. Objection 2 to this argument was that there *must* be a concept, JUSTICE, because it's commonly assumed that word meanings just are concepts. There is an English word, 'justice', and therefore its meaning must be the concept, JUSTICE. My response to this objection was to sketch a theory of word meaning that does not require that word meanings and concepts are the same things. Using Relevance Theory as an example, I showed how accounts of communicative success that hold that word meanings are concepts can be adapted so that this assumption is not in force. Furthermore, I suggested that any account of communication that holds that word meanings are concepts faces a serious challenge on the basis of the arguments I have been making throughout this thesis: if you think that the meaning of the word 'justice' is JUSTICE, then you must be able to provide an account of what properties JUSTICE has, or your theory is incomplete. If any of the analyses I have presented so far is along the right lines, then I think this will be a very difficult account to provide. On the other hand, if we abandon the assumption that word meanings and concepts are the same things, then we have found a way to avoid this problem altogether.

The main consequence of the theory I have sketched here is that the elements of a theory of concepts are not necessarily the same things as the elements of a theory of utterance interpretation. I think that, in the general case, the cognitive structures that result from interpreting utterances are not necessarily propositional thoughts built out of concepts (although such structures may be contingently generated in the interpretation of many utterances). There are also non-conceptual

cognitive structures involved in interpreting utterances, which are generated from information stored as a result of past usage episodes, and constrained by syntactic structure. These information contained in these past usage episodes of words and phrases is relatively unconstrained: from the Relevance Theory point of view, past usage information could take the form of any memory trace or any result of a cognitive process that was involved in producing a seemingly felicitous interpretation of a word on a previous occasion of use. Note however, that this information is not completely unconstrained. There is a reason that, in the absence of a strongly supportive context, we are unlikely to be successful if we use the word 'justice' to try and express thoughts about, say, the reproductive behaviour of cephalopods. The reason is that, as a matter of fact, in our language community people do not use the word 'justice' to try and express thoughts like these. In the next and final chapter, I will draw together all of the different arguments I have made throughout this thesis, and emphasise some positive implications from the preceding discussions. Hopefully, I will be able to convince you that abandoning the concrete-abstract distinction is no great loss to the study of concepts or cognition in general, and in fact has some clear benefits.

Chapter 9: Conclusions

In this chapter, I will draw together the various issues we have considered and emphasise some positive implications of the arguments I have been making. The goal of this thesis has been to convince you of two claims. The first claim is that the concrete-abstract distinction is not a useful psychological construct, and we should abandon it. The second claim is that the common assumption that words stand in a reliable correspondence with concepts is incorrect. Some words that we might think of as being 'abstract' do not pick out concepts, where concepts are construed as unitary cognitive resources. In the opening sections of this chapter, I will rehearse my arguments in support of these claims.

We started out in Chapter 2 by examining the notion of concreteness in psycholinguistic experiments. Concreteness has historically been an extremely important psycholinguistic variable for a number of reasons. The literature is generally presented as containing strong evidence for concreteness effects, whereby words with low concreteness scores exhibit processing differences relative to words with high concreteness scores. These findings are taken to be especially significant because of the common assumption that words stand in a reliable relationship with concepts. The thinking is that, if words with low concreteness scores exhibit processing differences relative to words with high concreteness scores, then the cause of these processing differences must ultimately reside in the conceptual system. The concrete-abstract distinction has, therefore, been used most extensively in order to investigate the conceptual system, and many theories and suggestions about its contents and format have been advanced on the basis of concreteness effects. I noted that, in list memory and EEG paradigms, concreteness effects seem relatively stable. Words with low concreteness scores are harder to remember than words with high concreteness scores, and N400 amplitudes to concrete words are larger than N400 amplitudes to abstract words.

However, in Chapter 2 we began to see that, in many other paradigms, evidence for concreteness effects is not as consistent as it first appears. In fMRI paradigms, results are highly variable and often in direct conflict. In lexical decision experiments, a decision latency advantage for both types of words has been repeatedly demonstrated, and at least one recent investigation concluded that concreteness is not actually correlated with lexical decision latencies at all (Brysbaert

et al., 2016). This is important because there have been many attempts to incorporate concreteness lexical decision data into theories of the conceptual system (Barber et al., 2013; Connell and Lynott, 2012; Kousta et al., 2011). However, if in reality there is no consistent effect of concreteness on lexical decision latencies, then this suggests that these attempts are in vain. Furthermore, as we saw in Chapters 2 and 6, even if there was a lexical decision advantage for one type of word over another, we would not be able to explain why this advantage exists. This is because the properties attributed to concrete and abstract concepts (e.g. multimodal information for concrete concepts; emotional-affective information for abstract concepts) do not obviously explain why there would be differences in lexical decision latencies across conditions. Finally, we saw in Chapter 2 that it has recently been claimed, on the basis of lexical decision data, that emotional affect is especially characteristic of (alleged) abstract concepts (Kousta et al., 2011; Vigliocco et al., 2013). I argue that this claim is incorrect because the concrete-abstract distinction does not separate emotion concepts from concepts of medium-sized objects in a principled way. We can tell the same causal story about the acquisition and 'representation' of emotion concepts as we can about object concepts, and so we have no grounds for calling emotion concepts abstract. So far then, we see that, although there do seem to be some consistent concreteness effects in list memory and EEG, in general the literature contains conflicting results, and the concrete-abstract distinction is sometimes drawn in an unprincipled way.

In Chapter 3, we changed tack somewhat, and we considered what philosophical and psychological theories have to say about concepts themselves. I suggested that, although there is clearly some disagreement about the nature of concepts, there is relatively widespread acceptance of a general model of concepts and thought. This model holds that words stand in a reliable correspondence with concepts, that thought is made out of parts, and that concepts are these parts of thought. I further suggested that, despite some important differences between them, both a Fodorian language of thought and a Barsalou-ian simulator theory are instantiations of this general model. Next, I presented some of my core arguments against the claim that words stand in a reliable correspondence with concepts. If it *were* true that this correspondence held, then that would imply that humans possess a JUSTICE concept (in virtue of knowing the word, 'justice'). However, I argue that neither a Fodorian language of thought nor a Barsalou-ian simulator theory gains anything from positing JUSTICE. There is no behaviour or cognitive process that JUSTICE plays an explanatory role in. Instead, both theories might do better if we

posit more specific mental representations that mediate specific cognitive processes and behaviours. Furthermore, the notion of JUSTICE raises deep problems no matter which kind of theory you want to endorse. A Fodorian cannot provide the individuation conditions that make JUSTICE the concept that it supposedly is. A Barsalou-ian cannot provide an account of how JUSTICE is acquired or what mental representations constitute it. However, both accounts fare better if we apply them to more specific concepts that might feature in situations that we use the word 'justice' to describe. As recent debates about the prospects of embodied cognition have shown (Barsalou, 2016; Goldinger et al., 2016; Mahon and Caramazza, 2008), this is an extremely important issue. *If* we could avoid the problems that JUSTICE raises, then our theories of concepts and cognition would be in a much better shape across the board. I think this is a substantial benefit of accepting my two primary claims and the implications that follow from them. The only change that needs to be made to a Fodorian language of thought or a Barsalou-ian simulator theory in order to accommodate the arguments I have tried to make is that we abandon the assumption that words stand in a reliable correspondence with concepts.

At the end of Chapter 3, we considered two important objections to the view that JUSTICE does not belong in a theory of concepts. Objection 1 was that the psycholinguistic literature has produced hundreds of concreteness effects, and explanations of these concreteness effects require it to be the case that words and concepts do stand in a reliable correspondence, such that reading a word 'activates' its corresponding concept. If this were true, it would suggest that the claim that JUSTICE does not belong in a theory of concepts must be false. Objection 2 was that theories of communication and psycholinguistic investigations of concepts frequently assume that word meanings and concepts are the same things. If this assumption is correct, then the claim that JUSTICE does not belong in a theory of concepts must be false, because the meaning of the word 'justice' must be the concept JUSTICE. I spent the rest of the thesis providing my responses to these two objections.

In Chapter 4, I started laying out my response to objection 1. My response to this objection is ultimately to argue that evidence for concreteness effects is just not as strong as it is believed to be. We saw that the concreteness measure itself has statistical properties that arguably invalidate it as a psycholinguistic tool. In the Brysbaert et al. (2013) concreteness norm database of 40,000 words, there is no word with a mean value in the middle of the scale for which that mean value is an accurate representation of participants' judgements. This means that concreteness ratings in the middle of the scale are essentially the product of noise. Worryingly, in a

survey of many different concreteness experiments, we saw that the 'abstract' stimuli featured in these experiments actually tended to come from the middle of scale, where the concreteness measure becomes uninterpretable. This on its own severely weakens the force of objection 1, because it means that so far psycholinguists have not actually been producing concreteness effects. Instead, they have been producing experimental differences between words that people agree about how to rate, and words that people disagree about how to rate. I also showed that if we wanted to only include words which 100% of the norming population claimed to know, and with standard deviations below 1, there are less than 300 abstract nouns to choose from, and many of these are morphologically complex oddities such as 'purposefulness'. This is also evidence against the utility of the concrete-abstract distinction. The concreteness measure is supposed to tap into a fundamental ontological distinction between two different kinds of cognitive entity. The problem is that on the truly abstract end of the scale there don't seem to be many representatives of one of these kinds. Finally, in Chapter 4 I showed that the midscale disagreement phenomenon does not just apply to Brysbaert et al.'s (2013) database. It applies to other scales that measure similar things to concreteness, such as imageability, for instance (Connell and Lynott, 2012; Cortese and Fugett, 2004; Schock et al., 2012), but it does *not* apply to scales that measure different things to concreteness, such as emotional valence (Warriner et al., 2013). This shows that the problem is specific to the concrete-abstract distinction as it is operationalised in psycholinguistics.

Now, even though I demonstrated that there are problems with how concreteness is operationalised in psycholinguistic experiments, and that in lexical decision and fMRI paradigms, results have not been consistent, it is still the case that statistically significant contrasts between concrete and abstract items have been obtained in list memory and EEG paradigms. So we might still think that objection 1 holds some weight. In Chapters 5 and 6, I report my own list memory and EEG experiments, which ensured that the contrast between concrete and abstract items was genuinely in force. Counterintuitively, under conditions that should have made concreteness effects more likely and stronger, no such effect was obtained in 3 out of 4 experiments (two of the memory experiments, and the EEG experiment). Indeed, experiments 1 and 4 produced evidence in favour of the null hypothesis. In the third list memory experiment, a small concreteness effect of 0.3 words on average was obtained. I take all of this to show that evidence for *concreteness* effects is not very strong. The reason that I stress the word 'concreteness' in the previous sentence is because, although we did obtain a statistically significant difference in experiment 3,

we must bear in mind that we obtained evidence in favour of the null hypothesis in two other experiments. It could be that some words with high concreteness values are easier to remember in list memory paradigms, but this could just as well be due to factors other than concreteness itself. The theories that predict a memory advantage for concrete words in list memory are, arguably, not compatible with the properties that items at the extreme ends of the concreteness scale actually have. For example, Walker and Hulme (1999) suggest that the advantage for concrete items is due to the fact that the representations that instantiate concrete concepts are richer than those that instantiate abstract concepts. However, in many cases, it seems much more attractive to suppose that the mental resources that we think of as being 'abstract' are instantiated by much 'richer' representations than those we think of as being concrete (again, consider what kinds of mental representations might be associated with 'doorknobs' versus those associated with 'spirituality').

My response to objection 1 is not to argue that there definitely are no such things as concreteness effects (from the point of view of the explanations that we have of these effects). Rather, the point is that the concreteness scales we use could not provide this evidence even if the effects are 'real'. This is because of the methodological problems with the scale which we explored in Chapter 4, and also because maximising a contrast along the scale did not produce unequivocal concreteness effects in the way that we would expect it to (as we saw in Chapters 5 and 6). Putting all of this together, I think we have gone some way to mitigating objection 1: a lot more work has to be done in order to show that concreteness effects are reliable enough to support this objection.

In Chapter 8, I set out my response to objection 2. This response consisted of a sketch of a theory of communication that does not require that word meanings and concepts are the same things. On this view, the cognitive structures generated during the interpretation of utterances are not *necessarily* conceptual. Instead, non-conceptual past-usage information constrains interpretations according to general pragmatic principles. I also showed how a popular theory of communication, Relevance Theory, could be adapted in this way. I also hope to have convinced you that, in many respects, the past-usage account of meaning does better than a conceptual account of meaning. This is because conceptual accounts of meaning that hold that the meaning of the word 'justice' is the concept JUSTICE cannot provide a satisfactory account of what properties JUSTICE has. On the non-conceptual view, this problem disappears, because the non-conceptual view is not committed to there being such a thing as JUSTICE in the first place.

I hope that, if we put all of this together, then we will have established the claims that I want to convince you of. The first claim is that the concrete-abstract distinction is not a useful psychological construct. Here are the reasons why I think you should endorse this claim. Evidence for concreteness effects is not strong. There are serious statistical and methodological problems with how concreteness is operationalised in psycholinguistic experiments. Theoretical explanations of concreteness effects suffer from deep problems, the most pressing of which is that supposedly 'abstract' stimuli do not plausibly have the properties that psycholinguists seem to assume they have. A recent trend to talk of abstract concepts as being specially characterised by emotional-affective information is not justified by any independent principles. We have no reason to suppose that emotion concepts are not concrete on any standard use of the term, and so this particular way of carving the concrete-abstract distinction falls apart. The second claim was that words and concepts do not stand in a reliable correspondence with each other. There is one main reason for you to endorse this claim: theories of concepts, cognition, and meaning do much better if the claim is true. Using JUSTICE as an example, I showed that some alleged concepts do not have any explanatory role in theories, and that JUSTICE is incapable of having the properties that theories of concepts and meaning want to ascribe to it. Ultimately, what guarantees the status of a theoretical posit is whether it helps us explain and/or understand some phenomenon. I think concepts *will* help us explain and understand human cognition. However, we have been making life very difficult for ourselves by assuming that all words of natural language naturally match up with the elements of these explanations. We have been trying to build theories of concepts that accommodate this assumption, without examining the assumption itself. If we abandon the assumption that words reliably pick out concepts, then we lose no explanatory power while dispensing with some deep problems for philosophical and psychological accounts of the conceptual system.

I want to end with a quote from Brysbaert et al. (2013, p. 909) themselves, and which I like very much: 'collecting a lot of information about a variable does not by itself make the variable more "real"'.

References

Allen, R., Hulme, C., 2006. Speech and language processing mechanisms in verbal serial recall☆. J. Mem. Lang. 55, 64–88. https://doi.org/10.1016/j.jml.2006.02.002

Allott Nicholas, Textor Mark, 2017. Lexical Modulation without Concepts. Dialectica 71, 399–424. https://doi.org/10.1111/1746-8361.12190

Altarriba, J., Bauer, L.M., Benvenuto, C., 1999. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. Behav. Res. Methods Instrum. Comput. J. Psychon. Soc. Inc 31, 578–602.

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., Treiman, R., 2007. The English Lexicon Project. Behav. Res. Methods 39, 445–459. https://doi.org/10.3758/BF03193014

Barber, H.A., Otten, L.J., Kousta, S.-T., Vigliocco, G., 2013. Concreteness in word processing: ERP and behavioral effects in a lexical decision task. Brain Lang. 125, 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Barsalou, L., Wiemer-Hastings, K., 2005. Situating abstract concepts, in: Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought. pp. 129–163. https://doi.org/10.1017/CBO9780511499968.007

Barsalou, L.W., 2017. Cognitively Plausible Theories of Concept Composition, in: Hampton, J.A., Winter, Y. (Eds.), Compositionality and Concepts in Linguistics and Psychology. Springer International Publishing, Cham, pp. 9–30. https://doi.org/10.1007/978-3-319-45977-6_2

Barsalou, L.W., 2016. On Staying Grounded and Avoiding Quixotic Dead Ends. Psychon. Bull. Rev. 23, 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Barsalou, L.W., 1999. Perceptual symbol systems. Behav. Brain Sci. 22, 577–609; discussion 610-660.

Barsalou, L.W., Santos, A., Simmons, W.K., Wilson, C.D., 2008. Language and Simulation in Conceptual Processing, in: Glenberg, A.M., Graesser, A.C. (Eds.), Symbols, Embodiment, and Meaning. Oxford University Press, Oxford, pp. 245–283.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-EFfects Models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Begg, I., 1972. Recall of meaningful phrases. J. Verbal Learn. Verbal Behav. 11, 431–439. https://doi.org/10.1016/S0022-5371(72)80024-0

Bentin, S., McCarthy, G., Wood, C.C., 1985. Event-related potentials, lexical decision and semantic priming. Electroencephalogr. Clin. Neurophysiol. 60, 343–355. https://doi.org/10.1016/0013-4694(85)90008-2

Binder, J.R., Westbury, C.F., McKiernan, K.A., Possing, E.T., Medler, D.A., 2005. Distinct brain systems for processing concrete and abstract concepts. J. Cogn. Neurosci. 17, 905–917.

Bonner, M.F., Vesely, L., Price, C., Anderson, C., Richmond, L., Farag, C., Avants, B., Grossman, M., 2009. REVERSAL OF THE CONCRETENESS EFFECT IN SEMANTIC DEMENTIA. Cogn. Neuropsychol. 26, 568–579. https://doi.org/10.1080/02643290903512305

Borghi, A.M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., Tummolini, L., 2017. The challenge of abstract concepts. Psychol. Bull. 143, 263–292. https://doi.org/10.1037/bul0000089

Brener, R., 1940. An experimental investigation of memory span. J. Exp. Psychol. 26, 467–482. https://doi.org/10.1037/h0061096
Brysbaert, M., Stevens, M., Mandera, P., Keuleers, E., 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. J. Exp. Psychol. Hum. Percept. Perform. 42, 441–458. https://doi.org/10.1037/xhp0000159
Brysbaert, M., Warriner, A.B., Kuperman, V., 2013. Concreteness ratings for 40 thousand generally known English word lemmas. Behav. Res. Methods 46, 904–911. https://doi.org/10.3758/s13428-013-0403-5
Burge, T., 1993. Concepts, Definitions, and Meaning. Metaphilosophy 24, 309–25.
Camp, E., 2015. Logical Concepts and Associative Characterizations, in: The Conceptual Mind, New Directions in the Study of Concepts. MIT Press, pp. 591–622.
Carston, R., 2012. Word meaning and concept expressed. Linguist. Rev. 29. https://doi.org/10.1515/tlr-2012-0022
Carston, R., 2010. Explicit Communication and 'Free' Pragmatic Enrichment, in: Explicit Communication, Palgrave Studies in Pragmatics, Language and Cognition. Palgrave Macmillan, London, pp. 217–285. https://doi.org/10.1057/9780230292352_14
Chumbley, J.I., Balota, D.A., 1984. A word's meaning affects the decision in lexical decision. Mem. Cognit. 12, 590–606. https://doi.org/10.3758/BF03213348
Coltheart, M., 1981. The MRC psycholinguistic database. Q. J. Exp. Psychol. Sect. A 33, 497–505. https://doi.org/10.1080/14640748108400805
Connell, L., 2018. What have labels ever done for us? The linguistic shortcut in conceptual processing. Lang. Cogn. Neurosci. 0, 1–11. https://doi.org/10.1080/23273798.2018.1471512
Connell, L., Lynott, D., 2012. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition 125, 452–465. https://doi.org/10.1016/j.cognition.2012.07.010
Cortese, M.J., Fugett, A., 2004. Imageability ratings for 3,000 monosyllabic words. Behav. Res. Methods Instrum. Comput. 36, 384–387. https://doi.org/10.3758/BF03195585
Crutch, S.J., Ridgway, G.R., 2012. On the semantic elements of abstract words. Cortex 48, 1376–1378. https://doi.org/10.1016/j.cortex.2012.05.010
Crutch, S.J., Warrington, E.K., 2005. Abstract and concrete concepts have structurally different representational frameworks. Brain J. Neurol. 128, 615–627. https://doi.org/10.1093/brain/awh349
de Groot, A.M.B., 1989. Representational Aspects of Word Imageability and Word Frequency as Assessed Through Word Association. J. Exp. Psychol. 15, 824–845.
de Vega, M., Glenberg, A., Graesser, A., 2008. Symbols and Embodiment: Debates on meaning and cognition. Oxford University Press.
Doest, L. ter, Semin, G., 2005. Retrieval contexts and the concreteness effect: Dissociations in memory for concrete and abstract words. Eur. J. Cogn. Psychol. 17, 859–881. https://doi.org/10.1080/09541440540000031
Dove, G., 2016. Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. Psychon. Bull. Rev. 23, 1109–1121. https://doi.org/10.3758/s13423-015-0825-4
Dove, G., 2014. Thinking in words: language as an embodied medium of thought. Top. Cogn. Sci. 6, 371–389. https://doi.org/10.1111/tops.12102
Dove, G., 2011. On the need for Embodied and Dis-Embodied Cognition. Front. Psychol. 1. https://doi.org/10.3389/fpsyg.2010.00242
Fiebach, C.J., Friederici, A.D., 2004. Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. Neuropsychologia 42, 62–70.

Fiebach, C.J., Friederici, A.D., Müller, K., von Cramon, D.Y., Hernandez, A.E., 2003. Distinct brain representations for early and late learned words. NeuroImage 19, 1627–1637.
Flom, P., Cassell, D., 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NESUG Conf. Proc.
Fodor, J., 1998. Concepts: Where Cognitive Science Went Wrong. Oxford University Press, Oxford.
Fodor, J.A., 2001. Language, Thought and Compositionality. Mind Lang. 16, 1–15. https://doi.org/10.1111/1468-0017.00153
Fodor, J.A., 1975. The Language of Thought. Harvard University Press.
Gee, N.R., Nelson, D.L., Krawczyk, D., 1999. Is the Concreteness Effect a Result of Underlying Network Interconnectivity? J. Mem. Lang. 40, 479–497. https://doi.org/10.1006/jmla.1998.2627
Geng, J., Schnur, T.T., 2015. The representation of concrete and abstract concepts: Categorical versus associative relationships. J. Exp. Psychol. Learn. Mem. Cogn. 41, 22–41. https://doi.org/10.1037/a0037430
Gernsbacher, M.A., 1984. Resolving 20 Years of Inconsistent Interactions Between Lexical Familiarity and Orthography, Concreteness, and Polysemy. J. Exp. Psychol. Gen. 113, 256–281.
Gleitman, L.R., 1989. The Structural Sources of Verb Meaning.
Gleitman, L.R., Cassidy, K., Nappa, R., Papafragou, A., Trueswell, J.C., 2005. Hard Words. Lang. Learn. Dev. 1, 23–64. https://doi.org/10.1207/s15473341lld0101_4
Goldinger, S.D., Papesh, M.H., Barnhart, A.S., Hansen, W.A., Hout, M.C., 2016. The poverty of embodied cognition. Psychon. Bull. Rev. 23, 959–978. https://doi.org/10.3758/s13423-015-0860-1
Hamilton, A.C., Coslett, H.B., 2008. Refractory Access Disorders and the Organization of Concrete and Abstract Semantics: Do they Differ? Neurocase 14, 131–140. https://doi.org/10.1080/13554790802032218
Harnad, S., 1990. The Symbol Grounding Problem [WWW Document]. Phys. D. URL http://cogprints.org/3106/ (accessed 7.25.16).
Holcomb, P.J., 1993. Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. Psychophysiology 30, 47–61. https://doi.org/10.1111/j.1469-8986.1993.tb03204.x
Holcomb, P.J., Kounios, J., Anderson, J.E., West, W.C., 1999. Dual-Coding, Context-Availability, and Concreteness Effects in Sentence Comprehension: An Electrophysiological Investigation. J. Exp. Psychol. 25, 721–742.
Huang, H.-W., Lee, C.-L., Federmeier, K.D., 2010. Imagine that! ERPs provide evidence for distinct hemispheric contributions to the processing of concrete and abstract concepts. NeuroImage 49, 1116–1123. https://doi.org/10.1016/j.neuroimage.2009.07.031
Hurvich, C.M., Tsai, C.-L., 1990. The Impact of Model Selection on Inference in Linear Regression. Am. Stat. 44, 214–217. https://doi.org/10.2307/2685338
Intons-Peterson, M.J., 1983. Imagery paradigms: how vulnerable are they to experimenters' expectations? J. Exp. Psychol. Hum. Percept. Perform. 9, 394–412.
Jager, B., Cleland, A.A., 2016. Polysemy Advantage with Abstract But Not Concrete Words. J. Psycholinguist. Res. 45, 143–156. https://doi.org/10.1007/s10936-014-9337-z
James, C.T., 1975. The role of semantic information in lexical decisions. J. Exp. Psychol. Hum. Percept. Perform. 1, 130–136. https://doi.org/10.1037/0096-1523.1.2.130
JASP Team, 2018. JASP.
Jessen, F., Heun, R., Erb, M., Granath, D.-O., Klose, U., Papassotiropoulos, A., Grodd, W., 2000. The Concreteness Effect: Evidence for Dual Coding and

Context Availability. Brain Lang. 74, 103–112. https://doi.org/10.1006/brln.2000.2340
Kiehl, K.A., Liddle, P.F., Smith, A.M., Mendrek, A., Forster, B.B., Hare, R.D., 1999. Neural pathways involved in the processing of concrete and abstract words. Hum. Brain Mapp. 7, 225–233.
Kounios, J., Holcomb, P.J., 1994. Concreteness Effects in Semantic Processing: ERP Evidence Supporting Dual-Coding Theory. J. Exp. Psychol. Learn. Mem. Cogn. 20, 804–823.
Kousta, S.-T., Vigliocco, G., Vinson, D.P., Andrews, M., Del Campo, E., 2011. The representation of abstract words: Why emotion matters. J. Exp. Psychol. Gen. 140, 14–34. https://doi.org/10.1037/a0021446
Kousta, S.-T., Vinson, D.P., Vigliocco, G., 2009. Emotion words, regardless of polarity, have a processing advantage over neutral words. Cognition 112, 473–481. https://doi.org/10.1016/j.cognition.2009.06.007
Kroll, J., Merves, J., 1985. Lexical Access for Concrete and Abstract Words. J. Exp. Psychol. Learn. Mem. Cogn. 12, 92–107. https://doi.org/10.1037/0278-7393.12.1.92
Kruschke, J.K., 2011. Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. Perspect. Psychol. Sci. 6, 299–312. https://doi.org/10.1177/1745691611406925
Kuperman, V., Stadthagen-Gonzalez, H., Brysbaert, M., 2012a. Age-of-acquisition ratings for 30,000 English words. Behav. Res. Methods 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4
Kuperman, V., Stadthagen-Gonzalez, H., Brysbaert, M., 2012b. Age-of-acquisition ratings for 30,000 English words. Behav. Res. Methods 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4
Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends Cogn. Sci. 4, 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6
Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161–163.
Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203–205. https://doi.org/10.1126/science.7350657
Kuznetsova, A., Brockhoff, P., Christensen, R., 2015. lmerTest.
Laming, D., 2003. Human Judgment: The Eye of the Beholder. Cengage Learning EMEA.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240. https://doi.org/10.1037/0033-295X.104.2.211
Lee, C., Federmeier, K.D., 2008. To watch, to see, and to differ: An event-related potential study of concreteness effects as a function of word class and lexical ambiguity. Brain Lang. 104, 145–158. https://doi.org/10.1016/j.bandl.2007.06.002
Lockwood, P.L., Apps, M.A.J., Roiser, J.P., Viding, E., 2015. Encoding of Vicarious Reward Prediction in Anterior Cingulate Cortex and Relationship with Trait Empathy. J. Neurosci. 35, 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015
Löhr, G., 2017. Abstract concepts, compositionality, and the contextualism-invariantism debate. Philos. Psychol. 1–22. https://doi.org/10.1080/09515089.2017.1296941
Louwerse, M.M., 2011. Symbol Interdependency in Symbolic and Embodied Cognition. Top. Cogn. Sci. 3, 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Luck, S.J., 2014. An Introduction to the Event-Related Potential Technique, second edition edition. ed. Bradford, Cambridge, Massachusetts.
Lynott, D., Connell, L., 2012. Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behav. Res. Methods 45, 516–526. https://doi.org/10.3758/s13428-012-0267-0
Machery, E., 2009. Doing without Concepts. Oxford University Press.
Maddock, R.J., Garrett, A.S., Buonocore, M.H., 2003. Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. Hum. Brain Mapp. 18, 30–41. https://doi.org/10.1002/hbm.10075
Mahon, B.Z., Caramazza, A., 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J. Physiol. Paris 102, 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004
Margolis, E., Laurence, S., 2015. The Conceptual Mind: New Directions in the Study of Concepts. MIT Press, Cambridge, Massachusetts.
Marschark, M., Hunt, R.R., 1989. A Reexamination of the Role of Imagery in Learning and Memory. J. Exp. Psychol. 15, 710–720.
McRae, K., Jones, M.N., 2013. Semantic Memory. Oxf. Handb. Cogn. Psychol. https://doi.org/10.1093/oxfordhb/9780195376746.013.0014
Miller, L.M., Roodenrys, S., 2009. The interaction of word frequency and concreteness in immediate serial recall. Mem. Cognit. 37, 850–865. https://doi.org/10.3758/MC.37.6.850
Morey, R., Rouder, J., Jamil, T., 2015. Computation of Bayes Factors for Common Designs.
Morr, P.E., 1981. Age of acquisition, imagery, recall, and the limitations of multiple-regression analysis. Mem. Cognit. 9, 277–282. https://doi.org/10.3758/BF03196961
Murphy, G., 2004. The Big Book of Concepts. MIT Press.
Nelson, D.L., McEvoy, C.L., Schreiber, T.A., 2004. The University of South Florida free association, rhyme, and word fragment norms. Behav. Res. Methods Instrum. Comput. 36, 402–407. https://doi.org/10.3758/BF03195588
Nelson, D.L., Schreiber, T.A., 1992. Word concreteness and word structure as independent determinants of recall. J. Mem. Lang. 31, 237–260. https://doi.org/10.1016/0749-596X(92)90013-N
Otfried Spreen, Rudolph Schulz, 1966. Parameters of abstraction, meaningfulness, and pronounciability for 329 nouns. J. Verbal Learn. Behav. 5, 459–468. https://doi.org/10.1016/s0022-5371(66)80061-0
Paivio, A., 2013. Dual coding theory, word abstractness, and emotion: a critical review of Kousta et al. (2011). J. Exp. Psychol. Gen. 142, 282–287. https://doi.org/10.1037/a0027004
Paivio, A., 1991. Dual coding theory: Retrospect and current status. Can. J. Psychol. Can. Psychol. 45, 255–287. https://doi.org/10.1037/h0084295
Paivio, A., 1986. Mental Representations. Oxford University Press, Incorporated.
Paivio, A., C, N., A, N., 1968. CONCRETENESS, IMAGERY, AND MEANINGFULNESS VALUES FOR 925 NOUNS. J. Exp. Psychol. 76, 1–25. https://doi.org/10.1037/h0025327
Paivio, A., Csapo, K., 1969. Concrete image and verbal memory codes. J. Exp. Psychol. 279–285.
Paivio, A., Khan, M., Begg, I., 2000. Concreteness and relational effects on recall of adjective-noun pairs. Can. J. Exp. Psychol. 54, 149–60.
Paivio, A., O'neill, B.J., 1970. Visual recognition thresholds and dimensions of word meaning. Percept. Psychophys. 8, 273–275. https://doi.org/10.3758/BF03212591
Paivio, A., Sadoski, M., 2011. Lexicons, Contexts, Events, and Images: Commentary on Elman (2009) from the Perspective of Dual Coding Theory. Cogn. Sci. 35, 198–209. https://doi.org/10.1111/j.1551-6709.2010.01146.x

Paivio, A., Walsh, M., Bons, T., 1994. Concreteness effects on Memory: When and Why? J. Exp. Psychol. Learn. Mem. Cogn. 20, 1196–1204.
Pexman, P.M., Hargreaves, I.S., Edwards, J.D., Henry, L.C., Goodyear, B.G., 2007. Neural Correlates of Concreteness in Semantic Categorization. J. Cogn. Neurosci. 19, 1407–1419. https://doi.org/10.1162/jocn.2007.19.8.1407
Plaut, D.C., Shallice, T., 1993. Deep dyslexia: A case study of connectionist neuropsychology. Cogn. Neuropsychol. 10, 377–500. https://doi.org/10.1080/02643299308253469
Pollock, L., 2017. Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. Behav. Res. Methods 1–19. https://doi.org/10.3758/s13428-017-0938-y
Prinz, J.J., 2004. Furnishing the Mind: Concepts and Their Perceptual Basis, New Ed edition. ed. A Bradford Book, Cambridge, Mass.
Putnam, H. (Ed.), 1979. Mathematics, Matter and Method, 2 edition. ed. Cambridge University Press, Cambridge ; New York.
Recanati, F., 2004. Literal Meaning. Cambridge University Press.
Rescorla, M., 2017. The Computational Theory of Mind, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.
Richards, L.G., 1976. Concreteness as a Variable in Word Recognition. Am. J. Psychol. 89, 707. https://doi.org/10.2307/1421468
Romani, C., McAlpine, S., Martin, R., 2008. Concreteness effects in different tasks: Implications for models of short-term memory. Q. J. Exp. Psychol. 61, 292–323. https://doi.org/10.1080/17470210601147747
Rubenstein, H., Garfield, L., Millikan, J.A., 1970. Homographic entries in the internal lexicon. J. Verbal Learn. Verbal Behav. 9, 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3
Ryan, T.P., 2008. Modern Regression Methods, 2 edition. ed. Wiley-Interscience, Hoboken, N.J.
Sabsevitz, D.S., Medler, D.A., Seidenberg, M., Binder, J.R., 2005. Modulation of the semantic system by word imageability. NeuroImage 27, 188–200. https://doi.org/10.1016/j.neuroimage.2005.04.012
Sadoski, M., Kealy, W.A., Goetz, E.T., Paivio, A., 1997. Concreteness and imagery effects in the written composition of definitions. J. Educ. Psychol. 89, 518–526. https://doi.org/10.1037/0022-0663.89.3.518
Schock, J., Cortese, M.J., Khanna, M.M., 2012. Imageability estimates for 3,000 disyllabic words. Behav. Res. Methods 44, 374–379. https://doi.org/10.3758/s13428-011-0162-0
Schwanenflugel, P.J., Akin, C., Luh, W.-M., 1992. Context availability and the recall of abstract and concrete words. Mem. Cognit. 20, 96–104. https://doi.org/10.3758/BF03208259
Schwanenflugel, P.J., Shoben, E.J., 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. J. Exp. Psychol. 9, 82–102.
Schwanenflugel, P.J., Stowe, R.W., 1989. Context Availability and the Processing of Abstract and Concrete Words in Sentences. Read. Res. Q. 24, 114–126. https://doi.org/10.2307/748013
Searle, J.R., 1980. Minds, brains, and programs. Behav. Brain Sci. 3, 417–424. https://doi.org/10.1017/S0140525X00005756
Skipper-Kallal, L.M., Mirman, D., Olson, I.R., 2015. Converging evidence from fMRI and aphasia that the left temporoparietal cortex has an essential role in representing abstract semantic knowledge. Cortex J. Devoted Study Nerv. Syst. Behav. 69, 104–120. https://doi.org/10.1016/j.cortex.2015.04.021

Sperber, D., Wilson, D., 1998. The Mapping Between the Mental and the Public Lexicon, in: Carruthers, P., Boucher, J. (Eds.), [Book Chapter]. Cambridge University Press, pp. 184–200.
Sperber, D., Wilson, D., 1995. Relevance: Communication and Cognition, 2 edition. ed. WB, Oxford ; Cambridge, MA.
Sperber, Wilson, D., 2015. Beyond speaker's meaning. Croat. J. Philos. XV, 117–149.
Tainturier, M.-J., Tamminen, J., Thierry, G., 2005. Age of acquisition modulates the amplitude of the P300 component in spoken word recognition. Neurosci. Lett. 379, 17–22. https://doi.org/10.1016/j.neulet.2004.12.038
Troche, J., Crutch, S., Reilly, J., 2014. Clustering, hierarchical organization, and the topography of abstract and concrete nouns. Front. Psychol. 5. https://doi.org/10.3389/fpsyg.2014.00360
Van Casteren, M., Davis, M.H., 2007. Match: a program to assist in matching the conditions of factorial experiments. Behav. Res. Methods 39, 973–978.
Van Petten, C., Kutas, M., 1990. Interactions between sentence context and word frequency in event-related brain potentials. Mem. Cognit. 18, 380–393. https://doi.org/10.3758/BF03197127
Vigliocco, G., Kousta, S., Vinson, D., Andrews, M., Del Campo, E., 2013. The Representation of Abstract Words: What Matters? Reply to Paivio's (2013) Comment on Kousta et al. (2011) [Editorial]. J. Exp. Psychol. 142, 288–291. https://doi.org/10.1037/a0028749
Vigliocco, G., Meteyard, L., Andrews, M., Kousta, S., 2009. Toward a theory of semantic representation. Lang. Cogn. 1, 219–247. https://doi.org/10.1515/LANGCOG.2009.011
Vigliocco, G., Vinson, D.P., Druks, J., Barber, H., Cappa, S.F., 2011. Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. Neurosci. Biobehav. Rev. 35, 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007
Wagenmakers, E., 2007. A practical solution to the pervasive problems of p values. Psychon. Bull. Rev. 14, 779–804.
Walker, I., Hulme, C., 1999. Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. J. Exp. Psychol. Learn. Mem. Cogn. 25, 1256–1271. https://doi.org/10.1037/0278-7393.25.5.1256
Warriner, A.B., Kuperman, V., Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav. Res. Methods 45, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x
Wechsler, S., 2015. Word Meaning and Syntax: Approaches to the Interface. Oxford University Press, New York, NY.
West, W.C., Holcomb, P.J., 2000. Imaginal, Semantic, and Surface-Level Processing of Concrete and Abstract Words: An Electrophysiological Investigation. J. Cogn. Neurosci. 12, 1024–1037. https://doi.org/10.1162/08989290051137558
West, W.C., Sitnikova, T., Holcomb, P.J., Caplan, D., Dale, A.M., 2001. Semantic processing of pictures of animals and tools: Event-related fMRI evidence on the organization of knowledge in the human brain. NeuroImage, Originally published as Volume 13, Number 6, Part 2 13, 760. https://doi.org/10.1016/S1053-8119(01)92102-0
Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behaviour? J. Anim. Ecol. 75, 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x
Wiemer-Hastings, K., Xu, X., 2005. Content Differences for Abstract and Concrete Concepts. Cogn. Sci. 29, 719–736. https://doi.org/10.1207/s15516709cog0000_33

Wilson, D., Carston, R., 2007. A Unitary Approach to Lexical Pragmatics: Relevance, Inference and Ad Hoc Concepts, in: Burton-Roberts, N. (Ed.), Pragmatics. Palgrave-Macmillan, p. 3.
Winnick, W.A., Kressel, K., 1965. Tachistoscopic recognition thresholds, paired-associate learning, and free recall as a function of abstractness-concreteness and word frequency. J. Exp. Psychol. 70, 163–168. https://doi.org/10.1037/h0022255
Wittig, M., Jensen, K., Tomasello, M., 2013. Five-year-olds understand fair as equal in a mini-ultimatum game. J. Exp. Child Psychol. 116, 324–337. https://doi.org/10.1016/j.jecp.2013.06.004

Appendix A

1. Experiment 1 stimuli

List	condition	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8
1	**disagree**	polling	dipstick	decade	centaur	exhaust	foreword	limbo	spender
2	**disagree**	physic	sequel	deacon	nettle	output	earshot	deadline	cackle
3	**disagree**	brethren	zenith	deluge	silence	lawsuit	theorist	polka	margin
4	**disagree**	nappy	degree	panic	bearings	legend	request	physics	prefect
5	**disagree**	sponsor	delta	dropper	phantom	egghead	rightness	aerial	eyesight
6	**disagree**	halter	brainwave	mankind	nightlife	surname	scrounger	tunic	omen
7	**disagree**	pariah	divorce	cosmos	sundries	purveyor	demon	crosswind	alias
8	**disagree**	grammar	conveyance	easement	blackball	woodland	giantess	weeknight	instant
9	**disagree**	tidbit	shallows	photon	plural	hallmark	grafting	sandman	nature
10	**disagree**	slipstream	audit	poorhouse	minute	rival	tribune	abyss	spectrum
11	**agree**	menace	bookie	tinting	flicker	rebound	squatter	tempo	pusher
12	**agree**	uprise	digest	tiling	region	charmer	joyride	outbreak	nutrient
13	**agree**	hubbub	matron	median	nuthouse	pullout	partner	distaste	refill
14	**agree**	burial	backwash	mover	career	event	footing	caper	peacetime
15	**agree**	jailbreak	torment	hazard	instinct	guru	downpour	richness	glucose
16	**agree**	bunting	rhythm	stalker	dullness	ascent	headache	gunpoint	welfare
17	**agree**	ringside	archduke	turmoil	shyness	posse	gangway	shipping	outreach
18	**agree**	sunburst	mishap	bumpkin	deceit	villain	bloodlust	misdeed	hunting
19	**agree**	diesel	roughhouse	attempt	whiner	viewpoint	freshness	stampede	leader
20	**agree**	semblance	havoc	broadside	dining	image	dissent	goner	culprit
21	**abstract**	setback	vagueness	spirit	notion	loyalty	esteem	phrasing	credence
22	**abstract**	charade	rapture	betrayal	logic	backlash	renown	letdown	affront
23	**abstract**	desire	mystique	intent	vantage	glory	nuance	unease	motive
24	**abstract**	amends	prestige	godsend	satire	leeway	wordplay	pretense	calmness
25	**abstract**	accord	whimsy	disdain	hardship	virtue	manner	regard	effect
26	**abstract**	freelance	mischief	respite	folly	pureness	repute	courage	meantime

27	**abstract**	merit	standpoint	future	allure	rapport	wisdom	prudence	insight
28	**abstract**	mistake	quantum	dogma	function	purpose	willpower	hearsay	meaning
29	**abstract**	patience	aspect	debut	fairness	pity	taboo	riddance	appeal
30	**abstract**	piety	finesse	foresight	longshot	loathing	stigma	concern	control
31	**concrete**	leaflet	roadhouse	artist	lighting	parsley	seabed	ironwork	lacrosse
32	**concrete**	clipper	pewter	cauldron	quarry	blockade	earwig	clubfoot	logbook
33	**concrete**	summit	breeches	abscess	foreman	award	entree	funnel	beacon
34	**concrete**	corset	template	pigment	fuchsia	urchin	ringworm	crewman	mansion
35	**concrete**	jester	gasket	sternum	backdrop	bouncer	chapel	resort	county
36	**concrete**	penthouse	fracture	entrails	vinyl	buckskin	tundra	barrier	plumbing
37	**concrete**	timepiece	methane	record	tiller	grindstone	merchant	shrapnel	duchess
38	**concrete**	quarter	bulkhead	sarong	tenant	chamber	canon	bailiff	machine
39	**concrete**	beaker	clinic	tango	clothing	amber	jackal	roulette	survey
40	**concrete**	spiral	marrow	billiard	bootlace	scabies	saffron	captain	product

2. Experiment 2 stimuli

Pair	Condition	Word1	Word2
1	**concrete**	cauldron	hike
2	**concrete**	footman	band
3	**concrete**	blazer	creature
4	**concrete**	rubble	liqueur
5	**concrete**	throttle	ulcer
6	**concrete**	ranch	gauntlet
7	**concrete**	cadet	concert
8	**concrete**	ledge	manor
9	**abstract**	betrayal	urge
10	**abstract**	revenge	foresight
11	**abstract**	godsend	risk
12	**abstract**	wisdom	psyche
13	**abstract**	hardship	malice
14	**abstract**	greed	riddance
15	**abstract**	loyalty	lenience
16	**abstract**	bliss	mercy
17	**midscale**	genius	royalty
18	**midscale**	foreground	district
19	**midscale**	gleam	patriot
20	**midscale**	view	approach
21	**midscale**	upstart	brawn
22	**midscale**	expanse	profit

23	**midscale**	asset	vortex
24	**midscale**	habit	encore

3. Experiment 3 stimuli

List	condition	word 1	word 2	word 3	word 4	word 5	word 6
1	concrete	pad	harpoon	stretcher	kennel	ulcer	aftershave
2	concrete	trachea	parsley	fuselage	rifleman	plaster	medallion
3	concrete	cedar	rubble	trinket	composer	liver	dormitory
4	concrete	scale	shipment	gladiator	guesthouse	morgue	marrow
5	concrete	vineyard	porcelain	cocktail	warship	advisor	slate
6	concrete	supervisor	infirmary	bouquet	manicure	bay	tomb
7	concrete	graphics	sage	smoothie	wildfire	prosecutor	sapphire
8	concrete	inspector	minefield	tourist	stub	horseradish	frostbite
9	concrete	guitarist	notch	gauntlet	orphanage	vegetation	bomber
10	concrete	greenhouse	sedative	museum	silicon	wreckage	accountant
11	concrete	incubator	lavender	surgeon	violinist	courtroom	embroidery
12	concrete	landlord	measles	dictator	pacemaker	minibus	plumber
13	concrete	newsletter	bodyguard	stockbroker	foliage	petroleum	liqueur
14	concrete	plantation	attorney	blockade	antibiotic	concert	currency
15	concrete	stroke	titanium	bile	sniper	massage	adhesive
16	abstract	urge	renown	patience	motive	malice	quandary
17	abstract	penance	belief	indulgence	reproach	version	fixation
18	abstract	mercy	glory	charade	aptitude	manner	formality
19	abstract	risk	psyche	rhetoric	foresight	fraud	regard
20	abstract	prudence	oblivion	hardship	mood	sarcasm	fate
21	abstract	extent	imposition	purpose	competence	luck	whim
22	abstract	willpower	bias	indecision	loyalty	seriousness	knowledge
23	abstract	involvement	existence	coincidence	ruse	principles	betrayal
24	abstract	detriment	subtlety	tradition	damnation	wisdom	fantasy
25	abstract	forgiveness	semantics	value	sanctity	godsend	discretion
26	abstract	eternity	politeness	concept	reasoning	anomaly	symbolism
27	abstract	suspicion	goodness	arrogance	mortality	chance	theory
28	abstract	precedent	privacy	likelihood	lunacy	oversight	revenge
29	abstract	affirmative	repentance	leniency	similarity	merit	expertise
30	abstract	wickedness	analogy	bliss	coercion	courage	avoidance
31	midscale	plot	molecule	mankind	format	swindle	motherland
32	midscale	hormone	reply	tarot	tribune	routine	pushover
33	midscale	delay	gossip	slumber	bandwagon	response	vigilante
34	midscale	zone	shallows	pinnacle	wavelength	grief	degree
35	midscale	envoy	character	fallout	clue	vacancy	tone
36	midscale	circulation	drunkenness	midsummer	doctorate	goal	hoax
37	midscale	cutthroat	rift	corporation	lawsuit	translation	sweetness

38	midscale	announcement	activist	process	slack	formation	whiplash
39	midscale	chronicle	monologue	overlap	motherhood	virus	penalty
40	midscale	exhaustion	delegate	magic	rebuttal	crackpot	diversion
41	midscale	entirety	ugliness	factor	ancestry	confidant	purgatory
42	midscale	engagement	accident	insomnia	regulator	utility	egghead
43	midscale	repellent	takeover	provision	dioxide	offence	thinker
44	midscale	equivalent	oracle	ignition	visibility	ransom	narrative
45	midscale	sense	extremity	content	lunatic	divorce	casualty

4. Experiment 4 stimuli

Sentence	Condition	
1	Concrete	The bay was a beautiful place.
2	Concrete	The plumber was coming to fix the car.
3	Concrete	The surgeon was preparing for the operation.
4	Concrete	The stretcher was brought out for the camel.
5	Concrete	The notch in the door had been repaired.
6	Concrete	The courtroom was cold and pink.
7	Concrete	Amy's newsletter was informative and well written.
8	Concrete	The sage was plucked fresh from Danielle's herb garden.
9	Concrete	The advisor of the defendant told him to lick the deal.
10	Concrete	The graphics were particularly impressive.
11	Concrete	The man's medallion was made of rice.
12	Concrete	The scale was broken and useless.
13	Concrete	The rubble was uncooked and difficult to climb.
14	Concrete	The marrow was cooked perfectly.
15	Concrete	The horseradish was invisible and tangy.
16	Concrete	The wildfire was decimating the forest.
17	Concrete	Luke's plantation of weeds had an excellent harvest.
18	Concrete	The dictator of the country was removed from power.
19	Concrete	Gabby's manicure was making her feel much better.
20	Concrete	Laura's guesthouse was highly rated on the wall.
21	Concrete	The infirmary was running out of space.
22	Concrete	The lavender was arguing in the summer sun.
23	Concrete	The man's plaster was coming off in the rain.
24	Concrete	The morgue was still singing for the body.
25	Concrete	David's liver was causing him problems.
26	Concrete	The guitarist of the hospital was renowned for his unusual technique.
27	Concrete	The bile was rising in Gary's throat.
28	Concrete	The petroleum was costly to paint.
29	Concrete	The woman's liqueur was upsetting her stomach.
30	Concrete	The man’s aftershave was cheap.
31	Concrete	The bodyguard of the president was awarded for his lunch.
32	Concrete	The orphanage of the asteroid needed donations urgently.
33	Concrete	The gauntlet was from the medieval period.

34	Concrete	The museum of fine eggs closed yesterday.
35	Concrete	The porcelain was filthy and chipped.
36	Concrete	The slate was covered in strange carrots.
37	Concrete	The woman’s landlord was strict but fair.
38	Concrete	The supervisor of the donkey told him to take a break.
39	Concrete	The stockbroker was afraid of losing his money.
40	Concrete	Martin's dormitory was on the second floor.
41	Concrete	Carol's embroidery was admired by the giraffe.
42	Concrete	Nigel's antibiotic was prescribed by the doctor.
43	Concrete	Paul's shipment of jokes was always late.
44	Concrete	The rifleman was cleaning his flower.
45	Concrete	The man’s cocktail was far too comfortable.
46	Concrete	The tourist was enjoying his time away.
47	Concrete	Billy's massage was crispy and relaxing.
48	Concrete	The inspector of the crime scene discovered the important evidence.
49	Concrete	The accountant of the firm was rewarded for malpractice.
50	Concrete	The greenhouse of exotic flowers was Vanessa's favourite place to play.
51	Concrete	The violinist was practising the sonata.
52	Concrete	The tomb of the ancient cat was never found.
53	Concrete	The woman’s ulcer was becoming infected.
54	Concrete	The foliage of the bricks obscured the view.
55	Concrete	The sniper on the oven lined up his shot.
56	Concrete	John's pad of paper was nearly volcanic.
57	Concrete	The parsley was the key ingredient in the meal.
58	Concrete	The composer of the window basked in the applause.
59	Concrete	The woman’s smoothie was cool and refreshing.
60	Concrete	Henry's kennel was full of hungry ants.
61	Concrete	The frostbite was starting to take its toll.
62	Concrete	The gladiator was victorious every time.
63	Concrete	Jane's minibus was packed full of frogs.
64	Concrete	The vegetation of the valley was lush and diverse.
65	Concrete	The woman’s trinket of paper sparkled in the light.
66	Concrete	Sarah's adhesive was not funny enough for the job.
67	Midscale	The goal of the initiative was to reduce cooking in schools.
68	Midscale	The character in the TV show was very popular.
69	Midscale	The woman’s divorce was chatting amicably.
70	Midscale	The woman’s plot was foiled by the investigator.
71	Midscale	The diversion was causing huge parties along the entire motorway.
72	Midscale	The provision was set out in the legal document.
73	Midscale	The woman’s doctorate was awarded on the moon.
74	Midscale	The hoax was planned out in incredible detail.
75	Midscale	The cutthroat was boarding the apple.
76	Midscale	The pushover was letting Jake get away with everything.
77	Midscale	The thinker was cartwheeling alone in his study.
78	Midscale	The sweetness of the purple girl was charming.
79	Midscale	The man’s reply was vague and unhelpful.

80	Midscale	Tammy's response was encouraging for the concerned giants.
81	Midscale	The circulation of the newspaper was very high.
82	Midscale	The pinnacle of the bookshelf was difficult to reach.
83	Midscale	The man’s whiplash was causing him problems at work.
84	Midscale	The envoy of the spoon was rewarded for his efficiency.
85	Midscale	The utility of the plan was questionable.
86	Midscale	The man’s drunkenness was ruining the rock.
87	Midscale	The hormone was activated by the treatment.
88	Midscale	The regulator of the industry fined the infants millions of dollars.
89	Midscale	The woman’s offence was struck off the record by the alien.
90	Midscale	The tone was annoying and played repeatedly.
91	Midscale	The activist was kicking an angry letter.
92	Midscale	The lawsuit was based on an obscure technicality.
93	Midscale	Molly's exhaustion was showing on her ear.
94	Midscale	Leo's translation was well received by the publisher.
95	Midscale	The vigilante was skipping from the law.
96	Midscale	The man’s corporation was filing for bankruptcy.
97	Midscale	The molecule was discovered in a playground.
98	Midscale	The lunatic was threatening the hostages.
99	Midscale	The delay of the train service pleased the travellers.
100	Midscale	The zone was declared off limits to the public.
101	Midscale	The process was drawn out over many centimetres.
102	Midscale	Robin's degree was in chemical engineering.
103	Midscale	The penalty was extremely delicious for the child.
104	Midscale	Danny's ransom was paid in full.
105	Midscale	The fallout of the dance lingered for weeks.
106	Midscale	The extremity of the injury soon became apparent.
107	Midscale	The visibility on the sofa was poor.
108	Midscale	The shallows were full of small fish.
109	Midscale	The chronicle of the kingdom was hated for its detail.
110	Midscale	The content of the speech was provocative.
111	Midscale	Jack's routine was starting to make him feel more energised.
112	Midscale	Anna's gossip was sparkly and hurtful.
113	Midscale	Ryan's engagement of seven months ended abruptly.
114	Midscale	The narrative was complex but the dogs praised it.
115	Midscale	The format of the TV show was innovative.
116	Midscale	The repellent was keeping the dinosaurs at bay.
117	Midscale	Julie's ancestry was a source of great pride.
118	Midscale	The vacancy was filled three centuries ago.
119	Midscale	Michael's slumber was interrupted by the doorbell.
120	Midscale	The ignition of the spatula was faulty.
121	Midscale	Oliver's accident was preventing him from training for the event.
122	Midscale	The clue was puzzling the detective.
123	Midscale	Phoebe's grief was finally coming to a roundabout.
124	Midscale	The man’s announcement was shocking to the nation.
125	Midscale	The virus was polite to all the animals in the area.

126	Midscale	The casualty of the incident was a young boy.
127	Midscale	Lucy's insomnia was a cause of celebration.
128	Midscale	The takeover of the business worried the shareholders.
129	Midscale	Charlie's monologue was spicy and boring.
130	Midscale	The rift was growing between the two political factions.
131	Midscale	The oracle was unsure of the colour of the battle.
132	Midscale	The delegate of the foreign country disagreed with the minister.
133	Abstract	The aptitude of the child was very impressive.
134	Abstract	The coercion of the termites was the subject of an
135	Abstract	The extent of the problem soon became clear to the
136	Abstract	Connor's forgiveness was easy to earn.
137	Abstract	The indecision was helping the team make progress.
138	Abstract	The woman’s malice was making her many friends.
139	Abstract	The oversight was costing the company lots of money.
140	Abstract	The quandary was resolved to everyone’s satisfaction.
141	Abstract	The similarity was easy for the shark to spot.
142	Abstract	The theory was ridiculed by the academic community.
143	Abstract	The wisdom of old woman was undisputed.
144	Abstract	The man’s arrogance was getting on Helen's nerves.
145	Abstract	The woman’s competence was plain to see so they fired
146	Abstract	Emily's fantasy was to become unemployed.
147	Abstract	The hardship was too much for the little boy to
148	Abstract	Colin's knowledge of the subject took him seconds to acquire.
149	Abstract	The mood was extremely tense in the ocean.
150	Abstract	Lauren's patience was running thin.
151	Abstract	The reasoning of the proof was flawless.
152	Abstract	The suspicion of the authorities was unfounded because Bob was
153	Abstract	The urge was difficult for Tim to resist.
154	Abstract	The charade was getting hard to maintain.
155	Abstract	The belief was widespread but incorrect.
156	Abstract	The concept was difficult for the snails to understand.
157	Abstract	Simon's fixation on meeting astronauts was bad for morale.
158	Abstract	The imposition was awkward for the diplomat.
159	Abstract	Ruby's leniency was despised by the children.
160	Abstract	The penance of the monk lasted forty days.
161	Abstract	The man’s politeness was insulting.
162	Abstract	Philip's repentance was unconvincing to his stern mother.
163	Abstract	The symbolism was lost on the cows.
164	Abstract	The version of the program was out of date.
165	Abstract	The purpose of Eric's actions was not obvious.
166	Abstract	The analogy was not very tasty.
167	Abstract	The betrayal was difficult for Diane to forgive.
168	Abstract	The discretion was greatly appreciated by the butler's cat.
169	Abstract	The formality was delightful but necessary.
170	Abstract	The man’s indulgence was starting to get expensive.
171	Abstract	Hugh's loyalty was unquestionably the worst thing about him.

172	Abstract	The precedent was ignored by the judge.
173	Abstract	The rhetoric of the campaign was divisive.
174	Abstract	The risk was too small for the explorers to take.
175	Abstract	The tradition was upheld for two minutes.
176	Abstract	The man’s whim was something he was coming to regret.
177	Abstract	The anomaly was detected by the incompetent footballer.
178	Abstract	Oscar's courage was failing him.
179	Abstract	The existence of forks on other planets is certain.
180	Abstract	The fraud was investigated by the burglars.
181	Abstract	The involvement of the police escalated the situation.
182	Abstract	Nina's luck was running quickly.
183	Abstract	The principles were impossible to argue with.
184	Abstract	The ruse was fooling everyone.
185	Abstract	Tom's sarcasm was welcomed by his tutor.
186	Abstract	The woman’s psyche was damaged by the relaxing day.
187	Abstract	Claudia's wickedness was revealed for the first time.
188	Abstract	The bias in the newspaper won it many awards.
189	Abstract	Chloe's expertise was reason she was sacked.
190	Abstract	The woman’s fate was yet to be decided.
191	Abstract	The glory of losing was good for the team's spirits.
192	Abstract	The lunacy of the idea made it popular with everyone.
193	Abstract	The motive of the criminal was clear.
194	Abstract	The prudence of the decision was widely recognised.
195	Abstract	The woman’s seriousness was making the situation more lighthearted.
196	Abstract	The subtlety of the argument made it hard to follow.
197	Abstract	Mark's willpower was fading quickly.
198	Abstract	The coincidence was particularly surprising for the butterfly.